

20.0 | 40.0 | 20.0

— 1805 — 1861 — 1901 — 1914 — 1939 — 1955 — 1989 — 2000

HISTORY By Numbers

^{2ND EDITION}
An Introduction to
Quantitative Approaches

PAT HUDSON and MINA ISHIZU

BLOOMSBURY

HISTORY BY NUMBERS

HISTORY BY NUMBERS

AN INTRODUCTION TO QUANTITATIVE APPROACHES

Second Edition

Pat Hudson and Mina Ishizu

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

B L O O M S B U R Y
LONDON • OXFORD • NEW YORK • NEW DELHI • SYDNEY

Bloomsbury Academic

An imprint of Bloomsbury Publishing Plc

50 Bedford Square
London
WC1B 3DP
UK

1385 Broadway
New York
NY 10018
USA

www.bloomsbury.com

BLOOMSBURY and the Diana logo are trademarks of Bloomsbury Publishing Plc

First edition published 2000

This second edition published 2017

© Pat Hudson and Mina Ishizu, 2017

Pat Hudson and Mina Ishizu have asserted their right under the Copyright, Designs and Patents Act, 1988, to be identified as Authors of this work.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

No responsibility for loss caused to any individual or organization acting on or refraining from action as a result of the material in this publication can be accepted by Bloomsbury or the authors.

British Library Cataloguing-in-Publication Data

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

ISBN: HB: 978-1-4742-9415-7
PB: 978-1-8496-6537-7
ePDF: 978-1-8496-6573-5
ePub: 978-1-8496-6572-8

Library of Congress Cataloging-in-Publication Data

Names: Hudson, Pat, 1948– | Ishizu, Mina.

Title: History by numbers : an introduction to quantitative approaches / Pat Hudson and Mina Ishizu.

Description: Second edition. | London; New York : Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc, 2016. | Includes bibliographical references and index.

Identifiers: LCCN 2016010853 (print) | LCCN 2016015081 (ebook) | ISBN 9781849665377 (paperback) | ISBN 9781474294157 (hardback) | ISBN 9781849665735 (PDF) | ISBN 9781849665728 (ePub) | ISBN 9781849665728 (epub) | ISBN 9781849665735 (epdf)

Subjects: LCSH: History—Statistical methods. | History—Research—Statistical methods. | BISAC: HISTORY / General.
Classification: LCC D16.17. H83 2016 (print) | LCC D16.17 (ebook) | DDC

907.2/1—dc23

LC record available at <http://lccn.loc.gov/2016010853>

Cover design: Clare Turner

Cover image: © Shutterstock

Typeset by RefineCatch Limited, Bungay, Suffolk

CONTENTS

List of Figures	ix
List of Tables	xii
Preface	xv
1 The Prospects and Pitfalls of History by Numbers	1
The growth of quantitative history	1
The computer revolution	3
From the elite to the masses	6
Descriptive statistics and inferential statistics	7
Time series and causal analysis	9
Sampling	10
Mathematical modelling	11
Quantification as a common language	12
Reliability of data	13
Statistical categories and comparability of data	16
Choice of technique, use and misuse	18
Analysis of results	19
Pitfalls of modelling	20
Conclusion	20
2 The Origins and Nature of Quantitative Thinking	23
Origins of the statistical movement in Britain	23
The meaning of statistics	25
Data display and collection	26
The Victorian statistical movement	29
Twentieth-century developments	33
Statistical theory	35
Positivism	38
Objectivity and prejudice	40
Conclusion	41
3 Arranging, Rearranging and Displaying Data	45
Types of data	45
Some definitions involved in regrouping data	50
The presentation of tables and figures	52
Initial questions about the data	53
Grouping data in a frequency distribution	54

Contents

Bar charts	62
Histograms	68
Pie charts and pyramid charts	71
Graphs: time series	74
Other graphs using independent and causal variables	78
Word clouds and similar figures in textual analysis	82
Cartograms	85
Conclusion	85
4 Summarizing Data: Averages and Distributions	87
The mean	87
The median	90
The mode	91
The geometric mean	91
Choice of average	92
Dispersion around the mean: standard deviation and variance	95
Dispersion around the mean: the coefficient of variation	98
Rank order dispersal measures	101
More examples of analysis of distributions from history	104
The normal distribution	112
Skewed distributions	113
Distributions with more than one mode	116
Conclusion	117
5 Time Series and Indices	129
Index numbers (indices)	130
The formation of indices	130
Composite indices	134
Construction of composite indices: some examples	136
Real indices	141
Time series: influences	144
Measures of trend: growth rates	146
Calculation of the trend line	149
Cyclical fluctuations and moving averages	153
Seasonal fluctuations	157
Irregular fluctuations	160
Vital statistics or vital variables	160
Conclusion	161
6 Relationships Between Variables	163
The null hypothesis	164
The contingency coefficient 'C'	165
The scatter diagram	169

Dummy variables	174
Correlation coefficient (<i>R</i>)	174
How strong is the relationship?	175
The form of the relationship	176
Lagged results	177
Spearman's rank correlation coefficient	180
The regression line	182
The coefficient of determination	184
Examples of correlation and regression analysis in history	185
Multiple regression models	189
Non-random error, autocorrelation and multicollinearity	194
Dealing with autocorrelation and multicollinearity	196
Conclusion	204
7 Sampling and Significance Testing	205
The purpose and procedures of sampling	206
Independent random sample	208
Systematic and stratified samples	209
Other sorts of samples	210
Sampling error	212
The normal distribution	212
The distribution of sample means	213
Estimation of the population mean and standard deviation from a sample	214
Samples and populations: some examples	214
Difference-of-means test	216
The significance of sample results	217
The significance of correlation and regression results	218
Conclusion	225
8 Economic History and Econometric History	241
Some definitions	242
The history of econometric history	244
Econometric history: first wave examples	246
Neoclassical model-building	251
Econometric tools and econometric history today	254
Comparative economic growth and new time series research	257
The models, the evidence, the reality	259
Conclusion	260
9 Historical Research, Computing and the Digital Revolution	263
Useful software types	263
The impact of advances in ICT on historical research and writing	275

Contents

Getting started with quantitative and qualitative historical research employing ICT	284
Research projects and ICT	286
Things to look out for when embarking upon a computer-aided piece of historical research (whether quantitative or not)	289
Conclusion	290
Glossary	295
Notes	307
Index	331

FIGURES

2.1	Frontispiece from William Playfair, <i>An enquiry into the permanent causes of decline and fall of powerful and wealthy nations</i> (1805)	27
2.2	Divorce, docks, dogs, drainage and drink: a page from <i>Mulhall's Dictionary</i> , 1884	30
2.3	Domestic animals of the world reduced to a common denominator (cattle): pictogram from <i>Mulhall's Dictionary</i> , 1884	31
3.1	Probability of wives and children receiving different types of legacy as an absolute gift, Consistory Court of Chester, 1800–1857	63
3.2	Subject matter of Gladstone's speeches in the House of Commons, by decade, 1833–1894	63
3.3	Plot analysis bar charts from <i>The Lady's Magazine</i> , 1793–1815	64
3.4	Distribution of McVitie travellers by income, 1920–1937	65
3.5	Bar charts of class structure, 1750–1961	66
3.6a	Facsimile of 1881 census, Glasgow, Govan and Galashiels	67
3.6b	Bar chart showing Glasgow population (municipal burgh) by place of birth, 1881	68
3.7	Income distribution of McVitie travellers, 1929 and 1936	69
3.8	The age distribution of slaves advertised for sale, Boston, 1720–1781	69
3.9	Frequency polygons (pointed and curved) of land tax payers in Sowerby, West Yorkshire, 1782	70
3.10	Frequency polygons of male and female mean annual mortality from tuberculosis at different ages, in selected periods	70
3.11	Pie chart showing proportions of UK re-exports by geographical or trading area, 1965	71
3.12	Pie charts showing the proportions of different types of fabric in the wardrobes of nobles, professionals, artisans and shopkeepers/wage earners in Paris in 1700	72
3.13	Landholdings in Sowerby by occupational group, 1782	73
3.14	Pyramid chart of the Victorian rich and poor	74
3.15	Profit rates in the worsted industry, 1840–1858	75
3.16	Impact of smallpox on average height, by age	76
3.17	Fan introduction and diffusion in coal mines through the Ruhr district, 1861–1900	76
3.18	Strength of Social Democratic Federation, Fabian Society, Socialist League and Independent Labour Party in London, 1880s to 1915	77

Figures

3.19 Number of feature films produced in UK, France and USA, 1911–1925	78
3.20 The relationship between marriage rates and detrended GDP per capita in the United States, 1887–1960 (First World War and Second World War excluded)	79
3.21 Distribution of female seasonal unemployment before and after enclosure in the counties of Bedfordshire, Cambridgeshire, Essex, Hertfordshire, Huntingdonshire, Norfolk and Suffolk	80
3.22 The dependency ratio, 1541–1871	81
3.23 Lorenz curves	82
3.24 Word cloud analysis of Frances Parkes, <i>Domestic Duties</i> (1829)	83
3.25 Cartograms showing age and origins of immigrants to Liverpool, 1851–1901	84
4.1 Pictogram of white-collar salaries in a firm in the 1950s	88
4.2 Height distribution of US passport applicants, 1830–1857	90
4.3 Wealth distribution in England	94
4.4 The geography of landholding inequality in Russia c. 1905	103
4.5 Real personal earnings quintiles for non-farm year-round workers by sex, 1940 and 2000	104
4.6 The effects of taxes and benefits on quintile groups of households, 1987	108
4.7 Median age at leaving home, United States, 1880–1990	110
4.8 The normal distribution	112
4.9 Social classes and genetic worth	113
4.10 Skewed distributions	114
4.11 Distribution of land tax payers, Sowerby, West Yorkshire, 1782	114
4.12 Age distribution of the Lancashire cotton industry workforce in 1833 and 1851	115
4.13 Distribution of sentence lengths, Portland Prison, 1849	116
4.14 Bi-modal distribution	116
4.15 Tri-modal distribution	116
5.1 Death rate per 10,000 worker years due to mining accidents in five European countries, 1851–1901	129
5.2 Indices of food prices, 1780–1870	134
5.3 Labourers' wages around the world	141
5.4 Respectability ratio for labourers: income/cost of respectable basket	144
5.5 Time series graph of burials and baptisms, St Martin in the Fields, London, 1554–1583	145
5.6 Trend line of baptisms, St Martin in the Fields, London, 1554–1583	150
5.7 Stylized graph to show the impact of selecting a correct (four-year) and an incorrect (five-year) moving average	156
5.8 Reported homicides (three-year moving average) and reported assaults per 100,000, New Zealand, 1878–1980	156

5.9	De-seasonalized movement of costs of provisions, Barrow workhouse, 1883–1886	158
5.10	Quarterly cash and credit sales of Jowitt and Company, wool staplers, 1832–1843	159
6.1	Scatter graph showing tea and sugar consumption, 1850–1865	171
6.2	Shares in farming, forestry and fishing and real income per person, 1946–2005	171
6.3	Railways per capita and exports per capita in Latin American economies, 1910–1914	172
6.4	Scatter graphs with different indications	173
6.5	Book production and the number of monasteries in Europe, sixth to fifteenth centuries	177
6.6	Scatter graph of rank of clothiers' ages at death against rank of value of assets	182
6.7	Relationship between ratio of harvest to spring wages and percentage of total acreage in arable land, 1836	183
6.8	Scatter graph of (a) wheat prices and number of disturbances; and (b) business cycle index and number of disturbances, 1810–1821	187
6.9	Impact of smallpox on average height, by age	189
6.10	(a) Homoscedastic errors; (b) non-homoscedastic errors	194
7.1	The normal distribution	213
9.1	Screen from Calverley land tax return, 1784	266
9.2	Screen from the land tax return for Calverley with Farsley, West Yorkshire, 1784	267
9.3	Page from Sowerby Baptism Register, 1730s	270
9.4	Land ownership and grants of waste, County Durham, 1100–1400	274
9.5	Family and social relationships in Austen	282

TABLES

3.1	Types of variables	45
3.2	Numbers of victims by gender of different thefts committed by female defendants in 1800	46
3.3	Average weekly expenditure for non-agricultural and agricultural working-class households, and for middle-class families with a head of household earning £250–350 per year, 1937–1938	47
3.4	Numbers in social classes, c. 1688	48
3.5	Fabrics in wardrobes of Parisians of different classes in 1700 and 1789	48
3.6	Colour analysis from bed hangings, 1660–1675, by room value range and room name	49
3.7	UK imports from, and exports and re-exports to, various regional groups and countries, 1965	51
3.8	The ratio of non-caps to caps and seats to spittoons in the best room of four pubs in Bolton, 24–28 January, unknown year in the 1930s	52
3.9	Potential violations of sumptuary law, 1327–1553	53
3.10	Return of convicts confined in Portland Prison, 1849	55
3.11	Simple frequency distribution of sentence lengths (unrelated to type of crime) of prisoners in Portland Prison, 1849	56
3.12	Grouped frequency distribution of sentence lengths (unrelated to type of crime) of prisoners in Portland Prison, 1849	56
3.13	Grouped frequency distribution of prisoners' origins as indicated by place of committal	57
3.14	Cumulative grouped frequency distribution of prisoners' ages, Portland Prison, 1849	58
3.15	Percentage cumulative grouped frequency distribution of prisoners' ages, Portland Prison, 1849	58
3.16	Analysis of hearth tax returns in the Yorkshire townships of Sowerby (including Soyland) and Calverley, 1664	59
3.17	Land tax payers, Sowerby, West Yorkshire, 1782	60
3.18	Grouped and percentage grouped frequency distribution of land tax payers in Sowerby, West Yorkshire, 1782	61
3.19	Manufacturers and tradesmen in Nottingham, 1783	61
4.1	Age at first marriage in Colyton, 1560–1837	93
4.2	Summary statistics of marriage in Cortona, 1415–1436	94
4.3	Distribution of defamation cases in three English courts, 1680–1687	95
4.4	Statistics relating to Table 4.3	97

4.5	Coefficients of variation of real wages, 1854–1939	99
4.6	Coefficients of variation of gross domestic product (GDP) per worker-hour, 1870–1938	100
4.7	Vehicle allocation in Soviet Russia during supply shocks and periods of normality	100
4.8	Number of archaeological remains found by twenty postgraduate assistants	102
4.9	Insured property of female and male slopsellers in London, 1777–1796	106
4.10	Average length of occupation leaseholds commencing in each decade, 1350–1409	107
4.11	Gini coefficients for the distribution of income at each stage of the tax–benefit system, 1975–1987	109
4.12	Estimated quartiles for number of years lived with one or both parents, United States, 1880–1990	111
5.1	Strikes in France and index of strikes, 1860–1870	131
5.2	Beer output, tobacco consumption and net income, 1925–1938	132
5.3	Indexed trends in output, labour force and output per worker, England, 1381–1700 and Great Britain, 1700–1851	133
5.4	Indices of wages of industrial and agricultural workers (selected years)	135
5.5	Indices of average money earnings per week and the formation of a composite index for selected years in the period 1780–1830	137
5.6	Output of key industrial sectors, England, 1270–1700 and Great Britain, 1700–1870	139
5.7	Components of an index of living costs, 1890–1900	140
5.8	Construction of an index of real wages, 1890–1900	142
5.9	Aggregate league attendance, gate receipts and average admission prices, 1927–1994	143
5.10	Annual baptisms, marriages and burials, St Martin in the Fields, London, 1554–1583	147
5.11	Calculation of the trend and of the detrended series of baptisms, St Martin in the Fields, London, 1554–1582	152
5.12	Mean heights (in inches) and the five-year moving average of mean heights (in inches) of English rural-born female convicts, aged 21–49 years, 1788–1819	155
5.13	Seasonally adjusted costs and residual fluctuations in provisions costs, Barrow workhouse, 1883–1886	159
6.1	Contingency table linking sentence lengths to types of crime, Portland Prison data, 1849	165
6.2	Contingency table linking level of education with family size	166
6.3	Businessmen: land ownership cross-tabulation, late nineteenth-century Britain	169
6.4	Per capita consumption of coffee, tea, sugar and tobacco, and an index of average real wages, 1850–1865	170

Tables

6.5	Lagged profit rates in the wool textile industry correlated with wage series for four firms, 1840–1858	178
6.6	Panel regressions on height in twelve European countries, 1720–1910	179
6.7	Age of death and value of moveable property of clothiers, with corresponding rankings, 1760s	181
6.8	Spearman's rank correlation coefficients relating to Swedish industrial growth, 1869–1912	182
6.9	Economic conditions and social tension in the early nineteenth century	185
6.10	Meta-analysis results of pock marks against height regression	189
6.11	Multiple regression analysis of London apprentices explaining their age at entering service	192
6.12	Conversion of a wage index to a series of first differences, 1790–1850	197
6.13	Aggregate grain price correlations, Soviet Union, April 1924 to October 1927	198
6.14	Estimates of the supply function of wolfram	200
6.15	The determinants of total social transfers as percentage of GDP, 1880–1930	201
6.16	Separation of trend and cyclical components from wheat prices for Winchester College, 1713–1718	203
7.1	Occupational status over time	219
7.2	Correlation of profit rates with export levels in the West Yorkshire wool textile industry for seven firms, 1822–1858	221
7.3	Correlation of profitability and wool prices in the West Yorkshire wool textile industry for the company T. & M. Bairstow, 1840–1858	222
7.4	The determinants of change in the agricultural wage in England and Wales, 1866–1912	222
7.5	The determinants of male migration rates from southern counties in Great Britain to six urban destinations, 1870s–1890s	223
7.6	Estimates of the dowry function, Cortona, 1415–1436	224
7.7	The sex ratio and the marriage rate of men, India 1931: <i>T</i> -statistics of the region effect	225
8.1	Annual returns on a 'prime field wench investment'	249
9.1	Details of a selection of computer files on Sowerby township in a relational database	268
9.2	Design for a family reconstitution database file of eighteenth-century baptisms	269

PREFACE

This book is a much revised version of *History by Numbers*, first published in 2000. All chapters have been redrafted with the addition of recent examples and exercises from historical practice. Many sections have been entirely rewritten to address the needs of students and colleagues tackling quantitative history in the twenty-first century. As with the first edition, no prior knowledge of statistics or of quantitative skills is assumed and techniques are presented in a straightforward, accessible manner. A short glossary is provided to assist those struggling with an entirely new ‘language’ but all technical terms are fully explained in the text as they are encountered. Exercises based upon published historical research, on a variety of periods and topics, are included after each group of chapters in order to test readers’ ability to apply critically the knowledge gained from the preceding pages. Chapter 9 has been substantially revised in order to guide historians, including students, embarking upon a research project using elementary quantitative methods with machine-readable data for the first time. The role of computers in speeding up and extending the scope of quantitative and qualitative enquiry is integrated in the body of the text with guidance about how best to employ them. Data analysis techniques are in all cases explained in relation to the role that computer software might usefully play in their application. An indicative list of electronically available datasets for historians and a list of other web-based resources and guides to assist with quantitative research are provided in a new electronic appendix, linked to the book and available on the Bloomsbury website at www.bloomsbury.com/history-by-numbers. We ask readers to assist the authors in future years (via email) in keeping this appendix updated, current, and useful for historians’ needs.

In the last half-century or so there has been a general extension of statistics as a language of enquiry in the social sciences and in public debate about economic and social conditions and policy across the globe. At the same time increasingly sophisticated statistical techniques have been applied to historical research in many fields: studies of crime, social protest, slavery, literacy, literary treatises, the composition and functioning of households, voting patterns, class structures, industrial output, population change to name just a few. Too often statistical data and analysis, whether relating to contemporary or historical evidence, is presented in a technical manner inaccessible to many readers. Research and conclusions in quantitative economic, social, political and cultural history may appear interesting or provocative but the reader is often faced with the necessity of taking the findings on trust. Fault lies with the readers of history as well as with practitioners. Even with simple quantitative expositions, many historians are content with ‘flicking through numerical matter as if it is not only distasteful but an unnecessary distraction from real historical analysis.¹ The mystification that can be created by quantitative work is

Preface

often accompanied by an abrogation of responsibility on the part of readers in confronting, questioning and appreciating the research. This limits intellectual debate and both the wider understanding and the necessary criticism, of quantitative analyses.

Quantitative work in history often provokes controversy and disaffection. This is partly to do with the reliability and suitability of surviving data, selected by historians for particular purposes, and partly to do with the manner in which the quantitative evidence is processed and manipulated. Too often, however, it is also because most historians do not understand basic and simple techniques of quantification. Many historians and most students of history are unable easily or accurately to challenge the authority of figures because some training in ‘history by numbers’ is essential to enter into a dialogue with those engaged in quantitative research. Historians do not necessarily want or need to become expert statisticians but they do need to be able to understand the language of statistical enquiry and to evaluate quantitative evidence and arguments that they encounter in their reading. Some will also wish to be able to undertake quantitative processing themselves for research or project work, especially if this involves primary sources and raw data that need to be simplified and summarized. At minimum, most historians recognize the need to understand those techniques of data display and analysis most commonly found in history as well as in wider social science, and in public life. Not to understand basic procedures is to be disenfranchised from significant areas of scholarly argument and debate, and from full participation in modern society.²

The main difficulty is that textbooks that seek to introduce scholars to quantification and statistical analysis in the social sciences, and specifically in history, tend to be too technical and inaccessible for most students.³ They throw people in at the deep end and spend little or no time explaining basic methods or how quantitative initiatives and ideas started and evolved: an obvious point of entry for a history scholar. The level of analysis and the technical nature of examples given in most statistics textbooks alienate history students from even the most basic quantification and encourage them to ignore quantitative evidence. This is a shame because some of the most important elements of quantitative analysis are very simple and straightforward. They involve only a common-sense reorganization and display of data. As with all types of historical evidence, quantitative data may be fragmentary, distorted or biased and historical skills will be required to evaluate and to analyse them. The statistical skills most commonly in use are in fact a great deal easier to acquire than the essential historical skills of evaluation and interpretation but they are not perceived as such. The association of quantitative history with dull and difficult abstract techniques rather than with the ‘normal’ processing and assessment of historical evidence promotes an unhealthy bifurcation in history and its research methods between quantitative and qualitative approaches, despite the fact that each shares many of the same methodological difficulties and potentialities.⁴ Many decades ago in an enjoyable little book, *How to Lie with Statistics*, Darrel Huff suggested that statistics are often used to ‘sensationalise, inflate, confuse, and oversimplify’ but, of course, the same could be said about words or prose: ‘without writers who use words with honesty and understanding, and readers who know what they mean, the result can only be semantic nonsense’⁵.

It is a good time to ensure that historians have a book of this kind to guide them but this is not because the popularity of quantitative history is growing. On the contrary, since the first edition of *History by Numbers* was published in 2000, the use of statistical analysis in history across most of its specialisms, has levelled off or been arrested. We except here a significant subset of research in economic history where, closely allied to economics, the subject has increasingly followed a quantitative path, seen especially in a current vogue for the comparative assessment and measurement of global economic growth.⁶ The more general decline in quantitative endeavour in history mirrors wider changes in the humanities and in many areas of social science, outside of economics. The decline of statistical analysis has resulted from a growing preference for descriptive, narrative and prose-based approaches and methods. Practitioners of quantitative methods are frequently accused of attempting to be more precise and free of subjectivity than the nature of social or historical evidence and the process of research can possibly allow. These sorts of criticisms are not new but they have grown in volume with the methodological debates surrounding post-modernism and with the linguistic or cultural ‘turn’ in history that arose in the 1980s and whose effects are still being felt. It is ironic that the linguistic turn has involved a heightened appreciation of the loaded and biased nature of vocabulary, language, terms and concepts: the very same appreciation that statisticians have always needed, when applying the categories and techniques of quantitative enquiry. Numbers and words, and their respective pitfalls, have more in common with each other than most of their users realize (and as we further explore in Chapter 1).⁷

The marked decline in the popularity and use of quantitative analysis has been accompanied by a decline in statistical training in higher education, in social science and humanities subjects in Britain (and, to a degree, in other Western academies) in recent decades. This has become a matter of concern and debate in government, in research bodies and in professional societies in Britain and elsewhere. It is feared that the decline will restrict the acquisition of skills vital to the economy and society, and to career development, both inside and outside of academia. This has in turn precipitated a flurry of activity. In 2008 the Economic and Social Research Council, in collaboration with the Higher Education Funding Council, launched an initiative to promote quantitative methods teaching in universities. A report was commissioned from Professor John MacInnes, *Proposals to support and improve the teaching of quantitative research methods at undergraduate level in the UK*, published in 2009, and a strategy to tackle the problem was developed.⁸ A year later, the Higher Education Academy, a body that promotes best practice in higher education teaching and learning published a report on numeracy skills amongst undergraduates with a title that conveyed wishful thinking: *Every Student Counts*.⁹ Further concern was expressed by the British Academy in 2012¹⁰ and a joint statement from learned societies in the arts and humanities, together with a response from the Royal Statistical Society soon followed.¹¹ A £15.5 million project by the Nuffield Foundation was set up in 2012 to attempt to rectify the problem and in 2013 HEFCE launched the £19.5 million Q-step strategy, a five-year programme to improve quantitative skills in social science at fifteen, mostly Russell Group, universities.¹² But

Preface

most of these strategies have had limited success thus far, partly because they focus upon high order statistical skills and ignore the need to start at ground zero, with an accessible and basic training adapted to the needs of those dedicated to research in particular fields, such as history.

Despite the turn towards cultural history and the general rejection of quantitative methodologies that this has entailed, descriptive statistics have maintained a strong place in some branches of history including business history, historical geography, demography, urban history, histories of consumption and crime and in some literary histories. Statistical analysis has remained in the ascendant across a swathe of economic history encouraged by the growing sophistication of modelling economic interactions and by new forms of historic national accounting. It also reigns supreme in areas of historical psephology. The problem is that because of the polarization between those who practise quantitative history and those who do not, quantitative historians end up talking largely to themselves. Fortunately, there are some recent promising signs of rapprochement between quantitative and qualitative studies in history. Production and consumption, trade, economic change, urbanization, migration and many other topics are being rethought as part of a boom in global history that is involving cultural, social and political historians as much as economic research. Cultural historians, for example, are finding it necessary to relate phenomena that they observe (for example, dress, customs, habits, behaviours, rituals, identities) to changes in material circumstances and to national and international exchanges. In tackling these subjects, quantitative evidence is inescapable: the traditional separation between qualitative and quantitative approaches is exposed to question and a wider range of historians is becoming more open to quantification.

New awareness and openness to quantification are also being encouraged by rapid advances in the digitization of primary sources and the production of electronically available datasets, many of which are available to scholars across the globe.¹³ This has been accompanied by a growing sophistication of software tools for both textual and numerical analysis that can now be done cheaply and effectively with only a micro-computer.¹⁴ However, historians need to have some understanding of the quantitative thinking and the methods of analysis that lie behind computer software packages. Too often computer-aided history is regarded as simply a matter of inputting data and applying the correct chain of software commands. This neglects the reasoning and logic behind quantitative methods and the need to be alive to the pitfalls and weaknesses, as well as the strengths, of both the techniques and the data. Computers have made the application of various analytical techniques so easy and accessible that procedures are often carried out with little understanding of their basis, their potential problems and the weaknesses and bias of the source materials. A straightforward discussion of basic techniques, their advantages and their problems, in the context of historical sources and historical enquiry, is much needed and this new edition of *History by Numbers* aims to fit that purpose.

We two authors have applied complementary skills and experience to producing this volume. Pat Hudson has fully revised her text from the first edition to which Mina Ishizu

has added technical comments and her cosmopolitan approach to understanding the needs of early career researchers and students in and beyond Britain. Both authors worked on the search for new illustrations, tables, graphs and exercises to accompany this edition; Ishizu taking a lead in this task. Our thanks go to the cohorts of economic and social history undergraduates and graduates who have participated in research methods and our other courses at the Universities of Liverpool, Cardiff and the London School of Economics between the 1970s and 2013. If the book proves useful to students of the future it will be partly a result of the constructive criticisms and feedback provided by their lively predecessors. We also thank colleagues of the two authors at Cardiff, at the London School of Economics, at Bristol, Sheffield, and elsewhere, who have long used *History by Numbers* and have lobbied for a new and updated version covering the same sorts of issues and skills. Colleagues struggling to teach quantitative methods to historians have been a source of support stressing the need for a short volume, accessible to all types of student, including those with an aversion to numbers. We must particularly mention Roger Lloyd Jones and Geoff Timmins whose work investigating the state of teaching and training in quantitative methods for historians, undertaken for the Higher Education Academy, was a spur to our efforts here, and Peter Howlett who kindly read and commented upon Chapter 9. Emma Goode and Emily Drewe at Bloomsbury have strongly supported the idea of a new edition from the outset and have been pivotal in seeing it through the press. Jim Oberly and an anonymous reviewer for Bloomsbury offered many helpful suggestions for which we are grateful. Above all we must thank Roger Middleton who has encouraged the production of a fully revised version of the book for several years, partly to coincide with the proposed new multi-volume *Abstract of British Historical Statistics*. To make full use of the large electronically available sets of key data that this project will produce, and of the 'Big Data' that is being generated by other digital initiatives across the world, students will need to keep an accessible guide to quantitative methods to hand.

Finally we must acknowledge and thank the many individuals, journals, organizations and publishers who have granted copyright for us to use tables, graphs and other figures as an illustration of the recent application of statistical and other quantitative techniques in published historical research. These include generous permission from Cambridge University Press for all the numerous book extracts, and for a table from *Continuity and Change*; and from Cambridge University Press together with the Economic History Association for all tables and figures from the *Journal of Economic History*. Wiley-Blackwell granted permission for the three extracts from *Historical Research* and Wiley-Blackwell together with the Economic History Society allowed permission for all figures and tables from the *Economic History Review*. Manchester University Press, Chicago University Press, the Pasold Research Fund Ltd (*Textile History*), Roderick Floud, the Royal Geographical Society and the Trustees of Mass Observation must be thanked for permission to use single extracts. Further tables and figures are included with permission from publishers, charged at commercial rates: Oxford University Press for extracts from the *Journal of Social History*, Taylor & Francis Ltd for extracts from *Social History*. We are particularly grateful to those on the list above who substantially

Preface

reduced or waived their copyright charges in order to further academic knowledge and training. All of the tables, graphs and figures in the text carry the required details of their original source.

History by Numbers aims to attract historians, at all stages of their training and research, to some basic techniques of quantitative history, not as a separate, marginal or suspect activity, nor as the be-all and end-all of being a historian, but as an essential tool and a necessary skill for everyone interested in the past. The volume emphasizes that quantitative skills are vital for critical engagement with other historians but that they need to be integrated with other approaches and used only where appropriate. Throughout the volume, we stress the ways in which quantitative history is compatible with, and complementary to, other sorts of historical analysis. We dedicate this volume to those of our students and colleagues who, like us, see no methodological distinction or incompatibility between quantitative and qualitative approaches to history and who champion the need for a combination of both.

CHAPTER 1

THE PROSPECTS AND PITFALLS OF HISTORY BY NUMBERS

In the hands of those who gather and use them, numerical data have the power to mesmerize and to control, to create order where none exists and to destroy an order or pattern by superimposing other divisions and categories. Numbers are not neutral: they are framed and defined by their creators, distorted and redefined by those responsible for their collection and reconstituted and reordered by those who select, display, use and analyse them for their own purposes.¹ In history, numbers play a key role in setting up debates and arguments and in creating periodizations and chronologies. Numbers are so central in history and in social science that they cannot be left to be gathered, manipulated and analysed by a restricted group of numerate academics. Too often the bulk of history readers and writers sit back and leave the numbers to someone else but this is a dangerous and self-limiting practice. All those interested in studying society whether in the past or the present need to take charge of quantitative data: to command it rather than to be the slave of a seeming authority of numbers emerging from documents or the writings of a small body of numerically inclined researchers.

In this chapter we consider the changing status and popularity of quantitative history over the past century. We look at its development and sophistication in some areas of history in the last few decades and the disenchantment with numbers that has occurred at the same time in other branches of our discipline. We discuss the various ways in which quantitative approaches and techniques can be of value in historical work but also the many drawbacks and difficulties involved in assembling and analysing historical data.² The chapter concludes by arguing that quantitative methods and numbers create both opportunities and potential pitfalls for historians but so too do historical approaches based on texts and words. Many of the problems are similar or identical.³

The growth of quantitative history

Quantification has long been a hallmark of much economic and social history and is the foundation of historical demography and historical sociology. Since the 1960s and 1970s it has also spread into other kinds of social, cultural and political history. One major factor encouraging this development has been a shift in the nature of history especially since the Second World War. Change from history based almost exclusively upon the lives of great men, elites, wars and diplomacy to histories of the mass of the population, the lives of ordinary families, women and children as well as men, the mass experience of economic growth and social change, made numerical analysis almost unavoidable.

History by Numbers

The French historian François Furet maintained that the integration of the subordinate classes into general history could only be accomplished through ‘number and anonymity’, by means of demography and sociology, ‘the quantitative study of past societies’.⁴ Largely for this reason he and others working within the Annales tradition in France, from the 1920s, encouraged a statistical approach to understanding the past. Their aim was to write ecologically-based histories in the round (*histoire totale*) of local and regional societies and communities over several centuries, which required the gathering and use of long-run statistical evidence of prices, output figures, population levels, wages and other similar indicators. This is often referred to as *histoire serielle* as it is based on the movement of series of vital variables relating to different sectors of the economy and different aspects of the quality of life. Emmanuel Le Roy Ladurie’s famous study of the *Peasants of Languedoc* was written very much from this perspective. His evidence of population, food prices and agricultural output, covering the sixteenth and seventeenth centuries, reflected the fluctuating fortunes of the rural labour force in the region, enabling him to observe the ‘immense respiration’ of the social structure.⁵ Meanwhile, social and political change in Britain and elsewhere in the twentieth century, the world wars and the development of welfare states influenced the ways in which historians viewed their subject and encouraged similar quantitative initiatives. These built upon work that had begun a century earlier, in many parts of Europe, when historians and commentators had begun their efforts to measure the impact of long-term economic and social change. In Britain the industrialization process was a major stimulus to charting the movement of economic variables, particularly agricultural and industrial output, prices, wages and trade figures. By the mid-twentieth century there was equal interest, from historians as well as economists, in estimating the movement of national income, expenditure and consumption, taxation and family budgets, as the role of the state in managing the economy expanded. (For more detail of these developments and the historiography see Chapter 2.)

The growing popularity of the social sciences, especially in Britain and the United States since the 1960s, further stimulated the wider use of quantitative analysis in history. The close relationship between economic history and economics, for example, encouraged quantitative techniques. In the United States in particular economic history was heavily influenced by econometric analysis: the application of economic theory via formal mathematical models of the operation of economic systems. Formalism (the algebraic expression and analysis of economic relationships of cause and effect) was taking a strong hold over economics at this time. By the 1960s attempts to understand the functioning of the economy or of business firms or sectors, using econometric tools, were finding their way into many branches of historical research, notably into analysis of the wisdom of technological choices made in the past. Many economic historians were converted to the idea that econometric techniques were revolutionizing the subject and injecting new certainties into historical debates.⁶ Those wedded to the claims of the so-called ‘New Economic History’ argued that definitive answers to many questions would now be possible and, at the extreme, it was anticipated that these new quantitative tools of analysis would reveal general laws of human behaviour which it had not been possible

to research before. A similar fever gripped sociology in the late 1950s and 1960s. Great stress was placed upon empirical research using mass survey data, large-scale interviewing and questionnaires. Sociologists were pioneering the development of statistical techniques, such as sampling theory to handle large bodies of data.⁷ Social and political historians were particularly influenced by this trend and a ‘new political history’ (largely concerned with voting patterns) and a ‘new urban history’ (mostly studies based on sampling from census data) emerged to match the new methods of sociological investigation and the ‘New Economic History’ in the 1970s and 1980s.

At the time Lawrence Stone complained that historians were becoming ‘statistical junkies’ and there was increasingly less contact and sympathy between those who were and those who were not immersed in the quantitative revolution.⁸ As some branches of economic history became generally more quantitative and less concerned with social, cultural and institutional matters, they crept closer to neoclassical economics in method and approach with its stylized facts and unquestioned, often anachronistic, implicit assumptions about the motivation and behaviour of individuals and groups. This inhibited any close alliance or integration of economic with social and cultural history, a characteristic that has dogged the development of economic history ever since, despite many attempts at rapprochement.⁹ As early as 1963 the President of the American Historical Association, with characteristic unconcern for gender sensibilities, warned colleagues about ‘worship at the shrine of that Bitch-goddess QUANTIFICATION’.¹⁰ By the 1970s and 1980s statistical approaches and their critics were reaching their height. Tony Judt likened quantitative historians to ‘clowns in regal purple’ suggesting that they had succumbed to a ‘delirium of statistical series’. What interested him was ‘not so much that historians cannot count but that they proclaim the need to do so’!¹¹ At the same time Liam Hudson remarked that ‘most social scientists who rely on … computers seem in practice to abandon their powers of reasoning’. Computer-aided quantitative work in history was seen to be causing an ‘atrophy of the critical faculties’.¹²

The computer revolution

In the 1960s and 1970s the computer revolution had begun in earnest, giving a major impetus to the spread of quantification in the social sciences and in history. Early computer use was both time-consuming and expensive. The need to employ an army of semi-skilled workers to feed in data on punched cards, to have access to expensive and slow mainframe machines and to employ specialists or to learn complex programming and computer languages, limited the use of computers amongst historians. Their major initial impact in history was in econometric history (where projects often shared the resources and expertise of economists working on allied subjects and often in the same departments) and in some branches of social history, which similarly drew on the experiences and skills accumulated in sociological studies. Despite their limited use, the potential for computers to extend quantification both within and beyond reasonable bounds in historical analysis was becoming recognized by the early 1970s. Richard Cobb,

History by Numbers

for example, warned of the degree to which the computer encouraged spurious and ad hoc manipulation of data merely to confirm already well-understood phenomena. He warned of ‘historians in white coats’: clinically involved with the manipulation of cold numerical data but detached from the heat of real historical research and historical sources:¹³

the computerisation of 516 urban riots . . . in France for the whole period 1815–1914. The end product will no doubt reveal some highly interesting patterns: that, for instance, market riots occurred on market days, on or near the market, that marriage riots take place after weddings, that funeral riots take place either outside the church or near the cemetery, that Saturday riots take place on Saturday evenings after the wineshops and bals have closed . . . that rent riots occur on rent days. . . . that religious riots, especially in towns or bourgs in which there exist two or more religious communities, favour Sundays, Catholic feast days, St. Bartholomew’s Day or the Passover. Perhaps we thought we knew already; but now we *really* know; we have a model. Riot has been tamed, dehumanized and scientified.¹⁴

Charles Tilly, whose work was the object of Cobb’s derision, felt forced to admit in an ensuing discussion that the scale and complexity of historical computer projects produce periods when the researcher is so preoccupied with problems of coding, file construction, statistical procedure, computer techniques and coordination of the whole effort that they practically lose contact with the people, events, places and times they are studying. He warned that:

In these days of the computer it is easy, tempting and relatively cheap to run large statistical analyses that are appropriate neither for the data at hand nor for the arguments that the investigator is really prepared to make. . . . The ease with which historical social scientists can run a hundred multiple regressions, carry out a large factor analysis, or compare every vote in a given legislature to every other one makes it easy to coax striking pseudo-results from almost any substantial collection of data.¹⁵

He recalled his quantitative training in the United States in the 1950s at the hands of Samuel Stouffer, the pioneer of social survey techniques:

An image of my early days as a graduate student sticks in my mind after almost thirty years. Half a dozen of us are standing around a clanking whirring machine in a harshly lighted basement room. There in the middle stirs sociologist Samuel Stouffer, talking fast, cigarette swinging from his mouth, ashes showering his vest. Stouffer grabs a deck of punched cards, shoves them into the hopper at one end of the machine, pushes a button, and watches the cards sort themselves into glass topped bins. He peers at the size of the various piles. Then he says ‘OK. Now let’s try breaking on religion’.¹⁶

These quotes illustrate a major problem that accompanied the first flush of enthusiasm and success in using computers to aid social and historical analysis and it has been a bugbear ever since.¹⁷ It was temptingly easy, once data was machine-readable, to run various tests of causal association between series until something that looked significant in a causal sense showed up. Thus the tail began to wag the dog because hypotheses based upon sound historical or sociological reasoning were being abandoned in favour of just waiting to see what the computer would ‘throw up’.

From the mid-1970s, computers rapidly became cheaper and more accessible with the microchip revolution and the innovation of personal and portable machines. More user-friendly software was also being developed. There was now no need to learn programming or a computer language. Instead, a range of software packages was available, including several specifically designed for the use of social scientists, and some for historians. Personal computers became more affordable with much larger memories enabling easy storage, retrieval and reordering of information as well as statistical manipulation. Projects that in the past would have needed a wealth of computer resources, staffing and a building to house the computer can now be undertaken by a lone scholar working at home or in an archive repository, with a laptop. This encouraged many historians, who had previously been hostile to computer use in their subject, to become zealous converts.

The software and hardware developments of the last five decades made many large-scale quantitative studies feasible, in financial and practical terms, for the first time.¹⁸ This is especially the case with projects that incorporate the searchable properties of digitized records. Many digitization and database projects have been completed since the 1970s offering a huge array of online, readily accessible, historical data. These developments present enormous opportunities for historians. Much online data is prose-based and software has been developed to aid the interpretation of literary style and content using basic statistical and computer recognition techniques. There is also a great deal of statistical information readily available in databases and spreadsheets. These derive from public as well as voluntary sources. In addition, in many countries publicly financed research comes with the obligation that any new statistical evidence should be stored in a publicly accessible (often a designated) repository. It is thus possible for a wide variety of historians to interrogate data produced by others for various new purposes. But to do this historians require a basic grasp of statistical techniques, as well as the historical knowledge critically to appreciate biases in the data that arise from many circumstances, not least the fact that they have been categorized and collected by someone else and for another purpose. It is more important today than in the past for historians of all kinds to acquire the range of skills necessary to make best use of the growing volume of digital data that is available.

Thus, widely available online historical data poses threats as well as opportunities for the practice of history. A major danger is the use of online data by historians with too little training in statistical techniques, who therefore do not know how to use the evidence to best advantage and can easily make basic errors of analysis. But a much larger danger is posed by the use of historical online data by researchers with too little knowledge of the reliability of the underlying sources and too few of the historian’s skills. Often, once

History by Numbers

databanks of historical statistical material have been generated they fall without their footnotes, and warnings about the uncertainties and difficulties of their creation, into the hands of those tempted by the ease with which they can be statistically manipulated. This will continue to happen as long as the bulk of members of the historical profession abrogate the responsibility of quantitative history, leaving it in the hands of those much less sensitive to the nuances of the data and much less aware of the pitfalls as well as the potentialities of their use. The widespread refusal of historians to engage with quantitative analysis also of course means that they are unable to critique published research that engages such techniques, from a position of knowledge or strength.

Early and continuing cavalier use of historical data, including digitized data, in quantitative history has given quantitative history in general a bad name in many quarters of the profession. The polarization between quantitatively and qualitatively oriented historians has also been endorsed in recent decades by the increasing popularity of a cultural history that generally concentrates upon analysing the sense of actions or symbols through detailed description of events or of texts and has little apparent need for quantification. We will see in Chapter 2 that much criticism of quantification in history is rooted in an opposition to the ‘scientific’ or positivistic approaches to history with which quantification is usually, but not always rightly, associated and which the new cultural history largely eschews. In the rest of this chapter we assess both the potentialities and the difficulties of quantification in history. This helps further to understand both the popularity of quantification within the historical profession and the widespread misgivings that it has also generated.

THE ADVANTAGES OF QUANTITATIVE HISTORY

From the elite to the masses

Many historians are attracted to broad statistical and numerical analysis because quantitative evidence is usually less elitist and more representative than qualitative data. It is perhaps surprising, given the greater opportunities that quantification presents for writing histories of the mass of the population, that so many historians of popular culture and society feel so negative about it. Personal papers and official records leave the historian with more information on elites than the working classes, on adult males than on women and children, on settled natives rather than migrant or ethnic minorities and on political and social activists rather than on the more passive majority of the population. Greater quantification can help to make best use of the documentation from the past particularly where that documentation deals with large numbers and with ordinary people.

For example, early statements that people married late in pre-industrial England were originally based on just a few pieces of evidence drawn from diaries and contemporary commentary of middle- and upper-class elites. Only a broadly-based quantitative study linking baptism and marriage data, from a large range of parishes, was able to provide

reliable evidence of gender-specific marriage ages and deviations from the average across society as a whole. This was a very important step in understanding the rise of fertility in many parts of Europe in the eighteenth century and in extending our knowledge from the small, mostly elite, section of the population whose lives are recorded in literary and official sources, to a wider and more representative group.¹⁹ Similarly in business history, much of our knowledge about capitalization and about entrepreneurial success or failure used to be based upon studies of a few large firms whose records had survived and were easily available but whose experiences were certainly atypical of the bulk of smaller or less successful firms. In recent decades major quantitative studies of business have appeared which have drawn upon insurance records, rate books, government surveys and bankruptcy figures. These are able to analyse the characteristics of a much wider variety of firms, including those that were small or short-lived.²⁰ Similarly textual analysis software can now aid in the statistical analysis of large quantities of digitized prose. Allowing historians and others to study the widespread use of certain words and phrases as an indicator of changing expression, thoughts, feelings, taste or fashion, the study of novels, diaries and letters en masse as historical sources has been revolutionized. Similarly printed sources such as newspapers, advice manuals and magazines can be analysed in new ways to indicate changing beliefs, cultural practices or patterns of consumption in a mass of the population, not just among an elite few. Once large datasets, that cover a range of experience, become recognized as a way of gaining a more representative and more accurate picture of general experiences, numerical analysis becomes essential.

Quantification in this context is associated with more representative and more accurate analysis than studies that are not willing to examine the characteristics of large numbers of cases. Such quantification can provide detail of patterns of experience of employment and unemployment, marriage, births and deaths, illness, crime, literacy, cultural practices and beliefs, tastes, material possessions, and fashion, for example, and can allow comparisons between different occupational or ethnic groups, regions, cities or nations. Qualitative social and cultural history, investigating what it was like for individual actors or families living these experiences on the ground, can then be placed properly in context.

Descriptive statistics and inferential statistics

Descriptive statistics refers to the reorganization or display of data in order to convey information more efficiently. Most history by numbers takes this form. The need to be both accurate and representative is facilitated by the clear display of data. Often, after just a little simple processing (into percentages or into an index, matrix or a histogram, pie chart or graph – all more fully explained in Chapter 3), data can be presented in a way that enhances our understanding of change or of particular circumstances in the past. A great deal of information can be relayed in summary form in this way. Where odd figures are merely quoted in a prose argument with phrases such as ‘larger than’ or ‘smaller than’

History by Numbers

or ‘rapidly growing’ or ‘steeply declining’, one is left in doubt about the available evidence, the representative nature of figures chosen for quotation and thus the validity of arguments. A table or figure representing the character of the data is very useful indeed. Elementary processing of figures to yield simple measures of average (or typical) experience and to give some idea of the range of variation in measurements over time or space is also valuable. The pitfalls as well as the benefits of numerical summary measures and display techniques are more fully investigated in Chapter 3. Whatever the criticisms of quantitative approaches to history, no historian should be unable to understand and to undertake simple statistical work of this kind. Providing one remains aware of the distortions that can be created by poor use of descriptive statistics, the ability to display data and to summarize its character accurately and concisely should be as essential a part of the armoury of a historian as the ability to read and to summarize a text.

Many historians and many more social scientists go well beyond simple descriptive statistics in their use of quantitative methods. Descriptive statistics are concerned with summarizing and describing a body of numerical information without suggesting that the observed patterns tell us anything beyond the information that is directly available. Inferential statistics, on the other hand, use the statistical data either to infer relationships between different sets of evidence or to extrapolate the data to other times, places or populations. This involves further processing, manipulating, modelling and testing of the data. Inferential statistics can be a powerful analytical tool if used with care but as the explanatory potential of statistical techniques increase so also do their pitfalls and dangers.

The foundation of inferential statistics is probability theory (more fully discussed in Chapter 7). Probability theory is important in quantitative analysis in all branches of study but particularly in those where it is impossible to repeat experiments hundreds of times in order to test the reliability of a hypothesis. In a subject like history (as with other social sciences) if we wish to test a hypothesis such as ‘the incidence of crime is closely related to unemployment levels’ we would have to rely upon the data of crime and unemployment rates which we have and which may well be limited. We are not able to experiment (as we would in the natural sciences) by adjusting levels of unemployment in order to see what happens to crime. The available data might well show that crime figures have a trend, or go up and down in cycles, which *appear* to shadow those of the unemployment rate but we will need to know if these patterns are very much more significant than patterns that might have occurred entirely by chance. Statistical measures of chance or probability are used to assess this. These measures of the significance of statistical findings can then be used as an indication (though no more than that) of real or historical significance. Inferential statistics can similarly assist in the extrapolation (or extension) of data trends and patterns beyond the figures available allowing one to suggest what missing values might have shown. In a similar way, the use of probability theory can indicate the likelihood that evidence from selected or surviving samples will reflect wider characteristics in an entire population.

Time series and causal analysis

As much historical research adopts a chronological perspective, it is not surprising that one of the most common quantitative methods used in history is time series analysis. Time series analysis is the study of the movement of measures, such as religious observance, literacy, wages, prices, exports, crime, the birth rate, wheat yields, over time. Measures such as these, that change with time, are always referred to as variables. As we shall see in Chapter 5, time series analysis involves the reconstruction and investigation of movements of a variable, or several variables, over time and usually comprises a mix of descriptive and inferential statistics. The construction of time series graphs, for example, enables one to see clearly the chronology of the rise and fall of outputs, prices, wages, strikes, murders, thefts, or whatever, revealing seasonal, or other periodic, fluctuations alongside longer term trends and tendencies. From such graphs and associated calculations, growth rates and rates of acceleration of growth and decline can be measured.

Time series analysis frequently requires the construction of indices. This involves converting original units into percentage measures, a simple process which is fully discussed in Chapter 5. Indices enable movements in several variables to be more clearly compared one with another even where their original units of measurement are different, for example comparing wheat yields (in bushels per acre) with wheat prices per bushel expressed in shillings. With the help of indices such variables as money wages can also be adjusted to take account of changes in prices to give a *real wage* series (a measure of the purchasing power of wages). Indices also enable one to calculate the overall movement of a number of separate variables in weighted combination (via a composite index). This is especially useful in calculating the movement of average wages from information on the wage movements of a spectrum of specific occupational groups, for understanding overall consumption of various goods based on sample figures from different social classes, or in calculating the overall output trends of the economy from evidence of the output of various separate sectors.

Time series and indices also allow the historian to consider how the movement of one variable over time may be related to another or to a whole series of others: for example how the readership of newspapers might depend upon the growth and social distribution of literacy rates; how prices over time might relate to the movement of wages; or how the incidence of riots may be related to levels of unemployment or changes in living standards. Techniques of statistical inference that are used to identify, isolate and measure the degree of association between two or more variables (whether they are in a time series or reflect geographical or social variation at one point in time) are very commonly used in quantitative social science and history. Examples include the relationship between party affiliation and voting behaviour; between type of crime and sentence length; between occupation and household structure; between wages and prices, between advertising expenditure and sales; between income and ownership of certain goods, between readership of newspapers or novels and their structure or design. Statistical techniques of correlation and regression can neither suggest nor prove that a

History by Numbers

relationship, let alone a causal relationship, exists between two or more variables but they can add support to a well-grounded hypothesis about causal connections and can indicate the strength or weakness of a possible relationship. These techniques, which are based upon probability measures, are discussed in Chapter 6.

Sampling

A further area of statistical analysis important in history is in the use of samples. Samples are generally used where there is too much data: where use of all the data available would be impractical, too costly or too time-consuming. Study of a small sample can yield accurate information about the population as a whole and informal sampling has long been used by historians of all kinds. For his oral history *The Edwardians* (1977), for example, Paul Thompson deliberately chose a representative sample of interviewees by region, class, occupation, sex. Similarly, Michael Anderson's study of household composition and change from the 1851 census aimed to get a representative 2 per cent sample by utilizing the census enumerators' books of one in every fifteen registration districts.²¹ Similarly Kevin Schürer and Matthew Woollard derived a 5 per cent random sample from the 1881 census for their research.²² A further recent example of sampling in historical research, and of debate surrounding the size and representative nature of the same, is provided in a debate about the impact of smallpox on the final height of adults in the nineteenth century.²³

Sampling theory (Chapter 7) alerts us to the dangers of choosing a sample which may not reflect the character of the wider population, instructs us in best practice regarding the selection of samples, enables us to predict the extent to which our sample results may be out of line compared with analysis of the whole population. Sampling theory, together with basic historical skills and judgement, also alert the historian to the biases inherent in 'samples' of data which just happen to have been recorded or to have survived. Surviving samples are the sort that historians often encounter and it is important to be able to judge whether the characteristics of such samples are likely to reflect wider experience. The analysis of sample results is again based upon inferential statistics and upon what we can draw from measures of probability (or chance) that a sample result will mirror wider characteristics or changes. These points are explored in Chapter 7. Any investigation that involves the consideration of sizeable amounts of data will potentially be aided by quantitative methods and approaches, usually with the aid of sampling and probability theory. This is particularly the case with data reflecting the movement of variables over long time periods and/or with research that questions the causes or the impact of certain changes or characteristics, where these are measurable.

One such field of historical enquiry that has felt the impact of quantification in recent years and that has expanded its role as a major branch of the discipline is demographic history. Under the influence of the Annales School, and of the growth of urban sociology, especially in the United States, groups of historians in most European countries and in Japan and America have established major research projects that consider the movement

of vital demographic variables (births, marriages and deaths). These studies are based on large samples and aim to reflect national movements, over long time periods. In Britain the work of the Cambridge Group for the History of Population and Social Structure has been important in this research. The aggregative analysis of parish register data has demonstrated that a decline in the age of marriage and a rise in the rate of marriage were more important causal variables than changes in mortality in accounting for early modern population growth in England. More intricate research on parish register data at local level, involving the reconstitution of families by linking the evidence of vital events, provides the statistical evidence for detailed demographic analysis of family size and structure, age of marriage and remarriage, population movements, age-specific mortality, and even enables calculation of the possible incidence of family limitation and breast feeding.²⁴ Demographic analysis has also developed using samples of census data in the study of household composition, migration, occupations and health while the history of medicine (incidence and causes of ill health and death, impact of hospitals and medical developments) has seen a major growth in the last decade using a mixture of statistical, biological and cultural historical approaches.²⁵ These fields have all benefited from the fact that early statisticians, reformers and civil servants were preoccupied with collecting figures relating to population, health and mortality, as we shall see in Chapter 2.

Mathematical modelling

Quantitative history was promoted by the popularity of mathematical modelling and the statistical testing of such models, which became popular in the 1960s and 1970s. Such modelling and testing remains an important branch within economic history in particular. In the last decade improved estimates of the long-term movement of many important national-level economic variables have been assembled: crop yields, rents, incomes, industrial outputs, Gross National Product (GNP), exports, imports, capital formation and other such variables across the globe in various ‘Big Data’ projects. See the electronic appendix to this book for the latest developments. The data collections of the Universities of Groningen (containing many comparative growth and development indices from around the world building upon refining the original estimates assembled by Angus Maddison) and Pittsburgh (which has taken on a similar mantle for an even wider range of comparative global data under the directorship of Patrick Manning).²⁶

Variables of this kind have been integrated into formal models of the functioning of the economy and have generated studies of the causes of movement in national income and output, living standards, wages and prices, industrial capital formation, factor productivity and so on.²⁷ The application of mathematical and statistical models of the behaviour of economic variables (econometrics) is increasingly used in economic history (econometric history), which is fully explored in Chapter 8.²⁸ The idea of some of these studies is to allow historians to assess the economic effect of a given policy or innovation by measuring it against the economic impact of what would have happened if the policy or innovation had not been implemented. In practice this involves measuring

History by Numbers

the outputs and costs in the real economy against their counterparts in a counterfactual (or hypothetical) case where a particular policy or innovation is absent. This approach has been applied to many issues. The most popular application has been to the impact of railway development in various countries but it has also been important in studies of American slavery, enclosure in Britain, and of late nineteenth-century economic policies in America.

Quantification as a common language

It can be argued that numbers, graphs and formulae should be regarded first and foremost as strategies of communication.²⁹ They are designed to convey information or results in a standardized form that can be understood across distances of culture, language and class. The language of quantification has been thrashed out over the past century or so as a highly structured and rule bound discourse that has enabled it to apply with little variation across the globe. The rules for collecting and manipulating numbers are widely shared and can be used to coordinate activities, to share research results and to agree policies. The growing use of quantification across many disciplines, including history, has thus been part of the globalization of research and of academic discourse more generally. This is a great strength of quantification but it is also the 'Achilles heel' in the eyes of its detractors. 'Precision' and ease of communication are often gained through the adoption of rigid categories, unquestioned assumptions and the loss of rich detail, all of which distort efforts to compare figures across time and space. Often, insufficient attention is paid to vagaries in the quality, reliability and transnational comparability of categorized data. Such difficulties are particularly prominent for example in transnational comparison of crime statistics where legal definitions of crimes and levels of detection vary greatly. Comparability issues across time and space are also a concern with widely used time series or comparative figures of Gross Domestic Product (GDP) per capita that are generally used to analyse the degree to which one nation may be ahead or behind another in terms of the level of economic development, or the pace of economic growth. The calculation of GDP figures is based upon widely varied taxation, census and other evidence over time and across economies. Like many comparative measures it is only accurate if it continues to measure more or less exactly the same thing. This brings us to the wide array of limitations that attend quantification.

LIMITATIONS OF QUANTITATIVE HISTORY

All approaches to historical study have their limitations and require the historian to be aware of them. The most severe limitations of quantitative history are however those it shares with all methods that use data from primary sources. Historical statistics, their display and manipulation, are similar to prose-based evidence in narrative or rhetorical argument: both are only ever as valid or reliable as the people or institutions that

produced them, on the one hand, and the historian herself who manipulates and judges them, on the other.

Reliability of data

A common pitfall in statistical research is that statistical methods are brought to bear upon data that are not sufficiently robust. Figures may be partial or incomplete (by design or by accident), wittingly or unwittingly distorted. Even where the degree of numerical error in an original source is limited, distortions can become magnified in statistical manipulation and this may then become crucial if the statistical analysis is the sole basis of a historical argument.

One must always bear in mind that the figures that pass down to us from the past were collected and assembled for purposes very different from those of the historian. A historian using official government papers, for instance, must always be aware of their general bias in favour of establishment and elite views of social problems, poverty, trade unions, strikes, commercial or agricultural depressions, for example. Historians using statistical evidence from government sources must also be alert to bias or selectivity in the way in which figures were selected and gathered. When using statistical data from any source, historians must remain aware of those who collected them, the purposes that they were geared to serve and the omissions and distortions that this will inevitably have introduced. Official values of exports and imports in the eighteenth and early nineteenth century, for example, were not altered to take account of price changes or innovations, and unofficial trade such as smuggling escaped record entirely. Thus the figures for the value of English international trade that the historian has to deal with for Britain before 1798 for exports and before 1854 for imports and re-exports can only provide a very rough idea of the growth in *quantity* of trade and relatively little at all about its changing value. They are thus of limited use for modern balance of payments assessments or as a measure of the value of output of particular industries.³⁰

Similarly, figures for unemployment for all periods are usually wildly inaccurate and distorted. This results from changing official definitions of unemployment, for example whether married women or students are included. They are also inaccurate because large numbers of people who are looking for work do not register as unemployed (especially those not eligible for benefits). The most significant bias in unemployment data is the omission of the bulk of unemployment of women and juveniles. Women's employment has traditionally been regarded as supplemental to household income. The unemployment of women is therefore seen as less important than that of adult males and it is less frequently registered and recorded. It was assumed that women and children would be reabsorbed into domestic subsistence activity during depressions in the nineteenth century. The shifting nature of education, the movement of the school leaving age and the changing treatment of student 'vacations' has resulted in poor measures of juvenile unemployment over the nineteenth and twentieth centuries, exacerbated by the short-term nature of training and apprenticeship initiatives.

History by Numbers

Restrictions on the benefit eligibility of unemployed married women and ‘students’ in twentieth-century Britain created a situation where many failed to register as unemployed (or employed) because there was no incentive to do so. This remains the case. Women and juveniles are thus often omitted from official figures and historians must make allowances for this.

Demographic evidence is also misleading largely because, as with a lot of sources, the figures provided are not those most appropriate to the historian’s needs. These have to be derived from what *is* recorded or known. Parish registers (our main source for population studies, marriage, mortality and disease in England before the introduction of Civil Registration in 1837) generally record evidence only of the parish populations who were of Anglican faith. Dissenters kept separate registers which have a poor survival record and non-Christians are missing from both Anglican and non-conformist documents. The numbers of those unrecorded often need to be estimated and added, a difficult task when non-conformity varied so widely across parishes and over time. In addition, parish registers provide baptism dates rather than birth dates and burial records rather than full information about death. Birth dates are more interesting and useful than baptism figures for the historian wishing to study population growth or fertility, not least because of deaths occurring between the two events. The birth/baptism interval is a source of major potential error especially when trying to estimate birth spacing or prenuptial conception rates. Similar problems occur in using burial records as mortality indicators. Burials do have the advantage of generally occurring within a week or so of the date of death but people were often transported back to their place of birth for burial so the record gets lost at the parish level.³¹

Census enumerators’ books for Britain and elsewhere are a commonly used historical source but have many pitfalls. Whilst reasonably accurate for male occupations, for example, they are hopelessly inaccurate in recording female work. Female occupational data was not regarded as a priority by those who designed the census nor by enumerators charged with collecting the evidence or giving advice to householders. Many women were also unwilling to reveal details of paid work because of the social stigma or tax implications. Low status, casual and intermittent employment was often not regarded as ‘work’ anyway. What should be regarded as ‘work’ was not well stipulated by the design of the census or the training of enumerators. In some of the late nineteenth-century censuses women working full-time in their family business were specifically not to be recorded as working. These sorts of issues surrounding women’s work affect most sources purporting to record such employment.³²

Taxation data has similar problems especially when used as a source to estimate population levels in the pre-census era. Records of hearth tax, window tax, land tax, poor rates refer to households and can only be used to estimate populations if a multiplier is applied. Gregory King estimated that the population of England and Wales in 1688 was 5.2 million on the basis of a multiplier of 4.5 of the numbers of households listed in hearth tax schedules. For three centuries nobody knew whether this multiplier had resulted in a reasonably accurate population estimate until Wrigley and Schofield’s broader quantitative work, based on parish registers, confirmed that King probably got it

about right. It is frequently necessary for the historian to use proxy variables: baptisms rather than births, *male* employment and unemployment figures as an indication of the employment structure of the economy as a whole, the ability to sign a register instead of real literacy levels. Sometimes historians use proxies which may be some way away from the variable which is really under consideration, for example, records of Easter communicants or the weight of candles burnt as an index of the ‘dechristianization’ of modern Italy and France respectively.³³

[Numbers seem] so central and seductive to the analytical mind, that if needed statistics are not to be found then the search for proxies gets underway. It is well regularly to recall that most of the commonly used historical statistics, at least before the mid nineteenth century, are in effect proxies. Numbers of manors held are made to act as surrogates for wealth; wage rates do duty for earnings; rents for profits. The heights of soldiers or marines are brought in to inform us about the standard of living.³⁴

This is a major problem with much quantitative work: decisions have to be made about the degree of reliability of proxy figures. ‘Guesstimated’ adjustments must often also be made to allow for omissions and distortions in the original data. Such adjustments are sometimes largely responsible for a major thesis such as with the timing of the late eighteenth-century baby boom in the work of Wrigley and Schofield.³⁵

All quantitative data should be subjected to historians’ judgements concerning reliability and accuracy, particularly where allowances have to be made to convert proxy variables to those required for the historian’s purpose. The questions should always be asked: who assembled the data and for what purpose? How were questionnaires or interview questions phrased and how may this have distorted the responses? How was the information sorted and reordered at the time of collection and what factors may have biased the survival as well as the recording of evidence? How accurate is quantitative evidence from the past, how attuned is it to the historian’s needs and how representative are those pieces of evidence and series of figures that have survived as an indication of wider trends and circumstances? There is also an important problem that often occurs with statistical measurements past or present: the process of measurement itself often causes shifts in behaviour that bias the results. If managers are told that their profit levels are to be the subject of measurement, there will be a strong incentive to optimize profitability by creative accounting in case this measurement is a prelude to new payment structures, bonuses or security of employment.

In economic and social affairs quantitative predictions and management by numbers often create inducements for business people, medical patients, tax payers and criminals (among others) to alter their behaviour in a way that undermines the numbers. That is, though the world described by social as well as natural scientists is partly a world of their own construction, they cannot make it however they choose.³⁶

History by Numbers

In the same way interviewees are likely to behave differently (and give different answers) in the dynamics of different interview situations, especially where the class, gender or ethnic character of the interviewer differs from that of the subject. The problem of distortion caused by the collection of data itself is not unique to quantitative work and closeness to the source carries no guarantee of greater precision.³⁷

Whether figures are sufficiently reliable is not an absolute question but depends upon the purposes for which they are required by the historian. If the figures are to be subjected to detailed statistical analysis or if they are to form the foundation of a thesis, they will need to be more reliable than if they are to be quoted simply as a supplement to an argument. Data from the past, whether quantitative or qualitative, will always be unreliable or unrepresentative to some degree. Whether a researcher accepts the degree of error that this injects into the narrative or analysis must depend on well-informed and well reasoned historical judgement.

Statistical categories and comparability of data

Statistical categories are initially devised by those who collect the data. These are usually private individuals, voluntary bodies or the state. They choose which figures to collect, and what categories to employ, to suit their own purposes. The collection of quantitative evidence has always been part of a strategy of individual improvement or advancement, social reform or state intervention so its categories are never simply descriptive. The discrete classifications into which varied and continuous information is often forced, frequently distorts the data in the interests of a larger purpose or project. This may be done consciously but is usually a problem even where efforts are made to avoid the distortion caused by the selection and classification of data. The big problem is how do we fit a rich and continuous variety of individuals into a manageable number of discrete categories for the purpose of data collection or later analysis? Where in any standard occupational classification, for example, should we fit a part-time but essentially retired taxidermist who is also a ski instructor and helps out in a charity shop? Reality is uncomfortably complex whilst statistical categories are necessarily gross simplifications. Sometimes the categories themselves should be questioned because they have been developed in such a way as to systematically minimize/exclude or maximize a particular group in the population, for example, an ethnic or religious minority. Historians need to be particularly aware of these issues not just as a guide to their own subjective bias but because they are also at the mercy of categories created by their predecessors and often have limited opportunity to rework or to adjust them. Categories are powerful because they often become entrenched and unquestioned even when they are misleading. If they are created by the state and are regarded as 'official' they are too often seen as unquestionable and representing real entities: they become 'reified'. Once a category of official statistics or classification gets fixed in the mind such as 'skilled white collar workers' and once it is used by commerce or the state as an object of policy, it can become a self-fulfilling entity as people themselves come to identify with the category.

The reliability and wisdom of classification criteria and thoughtful production of categories are important but so also is comparability, especially for studies of the same variables across nations, regions, localities, institutions, or over time. A series of figures is only accurate over time or across space if it continues to measure exactly the same thing. The longer the period covered, and/or the more diverse the geographical or cultural area, the less likely are series (whether wills, occupations, exports, sexual harassment cases or whatever) to be directly comparable. In coding death certificates two researchers in 1978 admitted that 'comparable statistics cannot be obtained if everyone does what he or she thinks is correct'.³⁸ Following rules may be necessary to achieve comparability of a series but it is often done at the expense of the richness of the evidence. The problem can become particularly serious where long-run or diverse statistical series are being used and with proxy variables. Crime figures are especially difficult to compare across time and space because of different legal systems and legal changes over time, different definitions of crime in different regions or countries, different responsibilities of various courts, as well as different or changing levels of policing and detection that can cause huge variation in the statistics for all or for particular offences. For example, contemporary police crackdowns on drink driving in the Christmas period create a major increase in records, for that crime, occurring in December that are largely caused only by increased detection and prosecution, rather than increased criminality. Periodic moral panics about particular crimes such as pickpocketing or paedophilia result in increased activity by police and courts that can create the appearance of a major crime wave in the recorded statistics.³⁹

Only a minor change in the basis of collection, measurement or assessment makes comparisons of statistical series over time very difficult. In the British census, for example, major changes occurred in occupational definitions and in the design of schedules from one census to another during the nineteenth and twentieth centuries as the census evolved to become a more efficient tool of social and economic management. Similarly in studies of living standards, changes in the composition of the workforce make the estimation of average money wages very difficult. Average real wage estimates (that is, estimates of the purchasing power of wages) are a further headache because it is necessary to take account of changes in consumption patterns over time (that affect the cost of living), as well as changing price levels. In studies of the movement of industrial output, changes in the sectoral composition of the economy, in the industrial mix and in the quality and types of outputs make the estimation of output growth problematic. In assessing the reliability and comparability of quantitative historical data, the skill of the historian is paramount. Assessment is likely to be based on familiarity with the institutional and administrative origin of the sources and with the historical context: knowledge of the likely missing and non-surviving data, how data were collected and for what purpose, how documents were conceived and how all these varied across time and space.

The same problems are often multiplied and certainly not eliminated when a historian chooses to use datasets collected by other historians. As we have already seen, the progress of computer technology in recent decades has boosted projects devoted to

History by Numbers

devising and to storing historical datasets so that once one set of historians has completed a piece of research the same data can be retained in a form useful for others to use for different purposes. Indeed, as we noted previously, this is often a condition where the collection and establishment of the database has been publicly funded. However, this is not as straightforward as it sounds because, even if the first historians have been very careful to record the pitfalls of the original data, the reasoning behind their categorizations and decisions, and how they have allowed for errors and omissions from the sources, these notes are not always carefully read or absorbed by those analysts who come to the data later and are further removed from the original source. As Clifford Geertz once famously wrote: 'what we call our data are really our own constructions of other people's constructions of what they and their compatriots are up to'.⁴⁰

Choice of technique, use and misuse

The historians' skills are also those most important in the choice of which statistical techniques to use and in decisions about how to use them. It is easy to get carried away with using statistical techniques and manipulation, especially when these are made relatively easy by digitization and computers, even when the data themselves are insufficiently representative or reliable to withstand such processing. Another common pitfall in the historian's use of statistics is manipulation of data that creates results or impressions that are misleading or false. At its simplest, this can occur, for example, in the choice of unit of measurement on the axes of a graph which can either magnify or reduce the appearance of trends and variations in the data. Choices about the display or analysis of data in distinct subperiods also have to be very carefully made as these can create misleading impressions of chronological change and indicate false discontinuities and turning points in data. For example, the timing and degree of discontinuity in English output and national income series for the eighteenth and nineteenth centuries depend very much on which subperiods are chosen for comparison.⁴¹ A number of other techniques are commonly used carelessly and in such a way as to create erroneous impressions or results; these include moving averages, correlation and regression analysis and sampling (more on all of these in later chapters).⁴²

Historians should always be on the lookout for deliberate as well as accidental misuse of techniques. This often occurs when statistics are placed in the service of some ideologically inspired thesis such as biological determinism. In his book, *The Mismeasure of Man*, Stephen J. Gould demonstrated the extent to which attempts to measure and predict human intelligence from late nineteenth-century craniometry (skull measurement) to sophisticated IQ investigations in the twentieth century, have been dangerous reflections of personal motives and of racial, class and gender prejudices. To take an example, the IQ-laden 11-plus examination for entry to grammar schools was soon shown to favour girls over boys at the age of 11. The pass level for boys was subsequently lowered in order to allow equal numbers of boys and girls to pass the exam.⁴³

Analysis of results

Finally, the skills of the historian are required in analysing the statistical results obtained from the manipulation of data as they do not speak for themselves. Estimating the reliability and significance of sample results can be a minefield as we shall see: measures that appear significant in a statistical sense may require a different assessment once the reliability of the original sources and the needs of a particular historical argument are borne in mind. In analysing the relationship between the movements of two variables over time a statistical result may suggest a close relationship and imply the possibility of a causal connection. But the statistical result may be quite accidental or spurious and only the historian can decide. There must be a good historical justification in seeking a statistical test for the existence of a relationship between two or more variables in the first place. We would be justified, for example, in suggesting that the price of cotton cloths on sale in Manchester during the Industrial Revolution would decline rapidly with increasing supplies deriving from the spread of mechanization. We might wish to explore the possible strength of the connection between the movement of price and output figures in the industry. Similarly, we might wish to investigate the links between income levels and the ownership of different consumer durables, such as cookers and fridges, in the 1950s and 1960s. But, a close statistical correlation or association between the movement of variables, however well founded our expectations of this may be, does not *prove* any sort of causality. It merely indicates that the historian's initial hypothesis *may* be justified and that it *may* be worth pursuing that avenue of enquiry. Statistical significance and historical significance are entirely separate phenomena.⁴⁴ The same is true of the difference between statistical and real significance in all other fields of study in both the natural and the social sciences. The historian and the reader must decide whether statistical results have any use or relevance to our understanding of the past. Statistics may point to the plausibility of possible explanations but they do not, in themselves, provide the answer to historical questions. Only the historian's interpretation of the results can do that.

Computers, by facilitating and speeding up the process of statistical manipulation have often resulted in bad as well as good statistical practice. It is now so easy to run a set of figures against several different series of causal variables that the temptation is there to do this and then to seek explanations for statistically significant relationships that appear. This is putting the cart before the horse: the hypothesis that there may be a causal connection between two or more variables (based on a sound historical argument or judgement) *must come first*, before the statistical analysis. And the hypothesis must be capable of falsification (of being proved incorrect).⁴⁵ If our statistical testing of a hypothesis does not yield supporting evidence we must be prepared to reject the hypothesis or to introduce another variable, *on the basis of another reasoned conjecture*, rather than to tinker around with the hypothesis just to make the data fit. Statistics may serve to reveal or to clarify a particular tendency but how we interpret that tendency, the significance of it, and the causal connections it may indicate, is a matter for seasoned historical judgement. The historical significance of results may also vary depending upon their use in an argument. If a minor and supportive piece of evidence is sought, the

History by Numbers

quantitative result may be accepted more readily than in a case where the quantitative analysis is the foundation stone of a whole argument or thesis.

Pitfalls of modelling

When we move from the use of descriptive statistics and simple statistical analysis of data to the use of quantitative techniques married to theoretical models of the functioning of variables, a whole new set of problems arise. In addition to those pitfalls already discussed, the model applied must also be scrutinized. If a model is applied to the past one must ask: does it embody a valid approximation of the behaviour of variables and actors at that time? Sometimes a neoclassical economic model is applied which has a supply/demand, free market bias, a limited number of variables derived from knowledge of the functioning of modern industrialized societies and a whole set of present-centred assumptions. (Most prominent of these is the rationality postulate which assumes that people and institutions will always act to maximize their profits or individual economic interests.) Such models generally assume that markets will move towards clearing at a certain level of price and hence of supply and demand. Sufficient information for buyers and sellers to act in such a way as to achieve this (in the medium term at least) is assumed. Neoclassical models also elevate those variables which are measurable to the most important in any analysis as these are the ones that the model is able to integrate. Non-measurable elements are given little attention.

Studies of manufacturing or transport innovations have commonly used econometric tools. But to evaluate innovations with any precision it is necessary to place a monetary value on every effect of the innovation. This is difficult in itself but also creates the problem of where to draw the line. Similarly in national income estimations using a national accounts framework (that relies on counting inputs and outputs and drawing a balance), there are major problems in constructing indices of economic growth for the eighteenth and nineteenth centuries. Surviving data from a sample of industries must be multiplied on the basis of what we know of the sectoral composition of the economy and of the relative growth rates of different sectors. There is a major problem of weighting the evidence that we have so that it can be used to estimate growth rates across the economy. And even if we manage to achieve a reasonably accurate estimate for this, it must be remembered that it is difficult to allow for changes in the weights over time and impossible to add on for the impact of innovations in the quality or design of goods, or for such important aspects of economic development as improvements in working hours and conditions, or the ways in which the benefits of economic growth are distributed.

Conclusion

In this chapter we have seen that, at the simplest level, quantification can bring to history the ability to summarize large bodies of data, to display such data effectively and to

express typical measures and values. In many cases quantification also result both in clearer specification and more rigorous testing of hypotheses about historical causation or relationships between variables. Statistical techniques may enable us to uncover important characteristics that are not apparent in the raw data and to confirm that relationships and patterns in the data are not there merely by chance. At the same time, vigilance is required in examining the source and reliability of data and the degree to which it may have been distorted by collection and prior processing. The danger of comparing statistical categories across time and space are legion and one should always therefore be on the lookout for 'results' of research that may merely emerge from variation in categories or statistical conventions across time and space. Above all, we must be aware of the dual sins of spurious attempts at statistical accuracy where the data is insufficiently robust to support this and the confusion of statistical measures of significance with a rounded evaluation of the historical importance of findings.

According to Peter Burke:

The introduction into historical discourse of large numbers of statistics has tended to polarise the profession into supporters and opponents. Both sides have tended to exaggerate the novelty posed by the use of figures. Statistics can be faked, but so can texts. Statistics are easy to misinterpret, but so are texts. Machine readable data are not user friendly, but the same goes for many manuscripts, written in illegible hands or on the verge of disintegration.⁴⁶

History is not an easy subject whether the historian chooses a quantitative or a qualitative approach or, as is often most appropriate, a mixture of the two. Neither should it be: the paramount requirement for all historians is to think hard about what they are doing and to be vigilant in interpreting evidence. Issues of the appropriateness of evidence for addressing the questions being posed – of reliability, representativeness and comparability/commensurability – must be uppermost in all kinds of historical research. In quantitative work, these issues must be examined and specified very precisely but once this is done the potential benefits of applying a range of statistical and quantitative techniques are great.

Further reading

Adyelotte, William O., Allan G. Bogue and Robert William Fogel (eds), *The Dimensions of Quantitative Research in History* (Princeton, NJ 1972).

Anderson, Ian 'History and computing' at: www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html (accessed 3 September 2015).

Fogel, Robert William and G. R. Elton, *Which Road to the Past: Two Views of History* (New Haven 1983).

Green, A. and K. Troup (eds), *The Houses of History: A Critical Reader in Twentieth Century History and Theory* (Manchester 1999), Chapter 6.

History by Numbers

- History Matters, at <http://historymatters.gmu.edu/mse/numbers/question1.html> (accessed 26 September 2015).
- Lee, C. H., *The Quantitative Approach to Economic History* (London 1977).
- Phillips, John L., *How to Think about Statistics* (6th edition, New York 2000).
- Rabb, T. K., ‘The development of quantification in historical research’, *Journal of Interdisciplinary History*, 13 (4), (1983), pp. 591–601.
- Tosh, J., *The Pursuit of History* (2nd edition, London 1991), Chapter 9.
- Tufte, E. R., *Visual Explanations: Images and Quantities, Evidence and Narrative* (Cheshire, CT 1997).

For some examples of interpreting quantitative historical source evidence see:

- Clubb, Jerome M., Erik W. Austin and Gordon W. Kirk, Jr, *The Process of Historical Inquiry: Everyday Lives of Working Americans* (New York 1989).
- Higgs, E. M., *Making Sense of the Census* (Essex 1989).
- Johnson, B. and S. Briscoe, *Measuring the Economy: A Guide to Understanding Official Statistics* (London 1995).
- Swierenga, Robert P. (ed.), *Quantification in American History: Theory and Research* (New York 1970).
- Taylor, Howard, ‘Rationing crime: the political economy of criminal statistics since the 1850s’, *Economic History Review*, 51 (3), (1998), pp. 569–590.
- Wrigley, E. A. (ed.), *Identifying People in the Past* (London 1973).

For the use of computers in historical work and the range of methods involved see:

- Cameron, Sonja and Sarah Richardson, *Using Computers in History* (London 2005).
- Greenstein, D. I., *A Historian's Guide to Computing* (Oxford 1994).
- Hudson, Pat, ‘A new history from below: computers and the maturing of local and regional history’, *Local Historian*, 25, (1995) reprinted in R. C. Richardson (ed.), *The Changing Face of English Local History* (Aldershot 2000), pp. 162–178.
- Lloyd-Jones, R. and M. J. Lewis, *Using Computers in History: A Practical Guide to Data Presentation, Analysis and the Internet* (London 1996).
- Mawdsley, E. and T. Munk, *Computing for Historians: An Introductory Guide* (Manchester 1993).

CHAPTER 2

THE ORIGINS AND NATURE OF QUANTITATIVE THINKING

The collection of numerical data and the roots of quantitative research in the study of society can be traced back many centuries. The pace quickened with the rise of centralized states in Europe and other parts of the world whilst the economic and social impact of industrialization created further economic and social imperatives for statistical accounting and analysis. The complexities of urbanized, industrialized societies with large state sectors, high taxation and a public commitment to macroeconomic management of various kinds has further resulted in massive data collection at local, national and international levels. Such central and local government statistics are a major source for historians alongside the mass of statistical material generated by private industrial and commercial concerns and voluntary bodies. The account that follows concentrates upon the history of quantification and of statistical methods and theories in Britain but the story has some direct parallels and overlaps with developments in other countries, especially France, Italy and the United States.¹

Studying the evolution of statistical approaches to society can tell us a great deal about the nature of data collected by individuals and by private and public bodies in the past. As with any other type of historical evidence, statistical information is always specified and gathered in relation to some particular commercial, legal, political, economic, moral or personal goal. The resultant figures with their assumptions, omissions, categories, biases and inconsistencies are the legacy with which historians grapple in their effort to interpret the past and to ask their own, very different, questions. Furthermore, just as the data collected in the past have been socially constructed by those who sought particular information for their own purposes or for purposes of state, so also the evolution of statistical methods and theories has been driven and moulded by particular social, political, moral or commercial goals. As we shall see, statistical thinking and statistical method have evolved within a social, economic and intellectual environment which conditioned their nature. Unless this is understood it is very difficult to comprehend both the opportunities and the controversies that surround quantification in the human sciences today.²

Origins of the statistical movement in Britain

Although the foundations of empirical social research in Britain are rightly associated with the Victorian statistical movement, the origin of quantitative study of economic, social and political problems can be identified much earlier. It should be seen as part of

History by Numbers

Enlightenment attempts to understand and to control society through rational, scientific enquiry and analysis. Some of the earliest work in this tradition in Britain is associated with John Graunt and William Petty in the 1650s and 1660s.³ In a series of papers William Petty argued the need for accounts of inhabitants in order to discover occupational structure and religious observance in England and Ireland. He also called for general registration of details of births, deaths and marriages in order to reveal age structures, religion, occupations and wealth in different parts of the kingdom. He established a new way of viewing society and of analysing social and political issues using *political arithmetic*. The foundation of political arithmetic was the idea that the prosperity and strength of the state rested on the number and condition of its subjects.⁴ Petty had been personal secretary to Thomas Hobbes in the 1640s and he may have been influenced in applying mechanical and statistical methods to social and political analysis by Hobbes' atheistic and authoritarian approach. He was also influenced in scientific method by the ideas of Francis Bacon.⁵ Petty's *political arithmetic* developed further during his time as Surveyor General in Ireland when he organized a detailed survey of County Down. Like his nineteenth-century successors, Petty saw the collection of statistical data and its analysis as an indispensable preliminary to a scientific understanding of the functioning of society and to the achievement of social and political reforms:

The method I take to do this is not yet very usual; for instead of using only comparative and superlative Words and intellectual Arguments, I have taken the course (as a specimen of the Political Arithmetick I have long aimed at) to express myself in terms of Number, Weight and measure; to use only Arguments of Sense, and to consider only such Causes as have visible Foundations in Nature; leaving those that depend upon the mutable minds, Opinions, Appetites and Passions of particular Men, to the Consideration of others.⁶

At much the same time, John Graunt was working on some of the earliest historical demography. His *Natural and Political Observations on the Bills of Mortality* published in 1662 analysed the London bills to consider urban and rural death rates, infant mortality rates, the excess of female births over deaths and the formation of life tables.⁷ Graunt, significantly, also discussed the reliability of social data, in particular whether different figures were sufficiently accurate to justify manipulation. He thus highlighted a consideration which remains central today in quantitative approaches to society.

Petty's followers in the statistical tradition in the late seventeenth and early eighteenth centuries, most notably Gregory King, concentrated on population and upon social structure although trade and public finance were growing concerns. King's estimate of the numbers and wealth of each rank of society in 1688 has become well known and used as a historical source, as in Table 3.4 (although it was not published in full until 1802). Charles Davenant and others concentrated at this time upon the statistics of trade with a view to efficient public finance and accounting. Davenant's definition of political arithmetic as 'the art of reasoning by figures upon things relating to government'⁸ has endured to show how literally statistical, that is 'state-thinking', this approach originally was and was to remain

for some time. There was a pragmatic force at work: the needs of an expanding and centralizing state bureaucracy for accurate estimates of wealth and revenues.

Many writers in the eighteenth century extolled the virtues of quantitative thinking and calculation especially in areas of public finance, including war finance, customs and excise, imports and exports.⁹ There were fifteen different estimates of national income alone between Petty's research and that of Patrick Colquhoun in the 1790s. In the same period progress was also made towards standardizing the myriad of local and customary weights and measures in England in the interests of internal trade.¹⁰ This can be seen as vital in establishing the conditions of an exchange society with common understandings about the weights and qualities of goods being transacted. It is an illustration of one of the main purposes of the evolution of commonly understood statistical measures and methods: that they function as a form of common discourse across geographical space and differing local cultures.

Improvements in the natural sciences, in the measurement of longitude, time, pressure and temperature, were matched by calendar reform and the Ordnance Survey (from 1791).¹¹ The eighteenth century also saw major progress in developing the principles of life insurance using life tables which had been introduced in Graunt's *Natural and Political Observations* and developed in 1686 by Edmund Halley (of comet fame).¹² Additional work done for the Scottish Ministers' Widows Fund, founded in 1744 by Alexander Webster and Robert Wallace; the development of the mathematical theories of chance by de Moivre and Simpson in the 1760s; and progress made under the auspices of the Equitable Assurances for Lives and Survivorships (founded 1762), were particularly significant.¹³ Probability theory was being advanced in the physical sciences as well as in relation to the insurance industry. Pierre Simon Laplace built on the work of De Moivre and both he and the Reverend Thomas Bayes (apparently independently) contributed to modern probability theory by developing the a posteriori technique. This enabled the prediction of events in the future, or the inference of causation from a record of past occurrences.¹⁴ Interestingly, the Gambling Act of 1774, which attempted to define insurance and chance as opposites can be seen to have marked an important transition in the acceptance of actuarial 'certainties'. Life insurance was no longer viewed as a gamble, nor length of life seen as random: both were coming to be seen as calculable mathematical probabilities. The insurance of sickness, where it was much more difficult to calculate the odds, remained much more risky and more a matter for subjective assessments, usually based on interviews, until well into the nineteenth century. Chance was being tamed but the general drive for objective rules to replace subjective evaluations in life assurance and in other fields (an important element in the rise of statistical thought and method) was a slow process.¹⁵

The meaning of statistics

The word statistics probably first entered the English language in 1770 with W. Hooper's translation from the German of a book by J. F. von Bielfeld.¹⁶ The meaning of statistics here was linked to the notion of statesmanship and defined as 'the science which

History by Numbers

teaches us the political arrangement of all the modern states of the known world.¹⁷ This *Staatenkunde* notion can be traced back to Aristotle but it became a serious, albeit contested, academic discipline in Germany in the eighteenth century.¹⁸ The word statistics was firmly situated in the English language by Sir John Sinclair. The first volume of his massive survey *A Statistical Account of Scotland*, appeared in 1791. Sinclair defined statistical enquiries as those 'respecting the population, political circumstances and production of a country and other matters of State' and he was so convinced of their importance that he advocated sending statistical 'missionaries' around the country.¹⁹

By the end of the eighteenth century a shift had started to occur in the definition of statistics in Britain away from the idea of ordered facts (both numerical and non-numerical) which would reveal the condition of the state, in the direction of more narrowly defined quantitative evidence, but change was slow. The 1797 edition of the *Encyclopedie Britannica* still defined statistics as 'a word lately introduced to express a view or survey of any kingdom, county or parish'²⁰ but the association of statistics with description or display of *numerical* (as opposed to other sorts of) data was developing. Statistics in France was identified by Charles Dupin as numerical information about society as early as 1820.²¹ And in his *The Statistical Breviary* of 1801, William Playfair implied that statistical works should be limited to quantitative data.²² Yet as late as 1842 J. R. McCulloch rejected the idea 'that everything in statistics may be estimated in figures'.²³ It was not until the twentieth century that statistics came specifically to mean the arrangement and manipulation of purely quantitative evidence: a development that had involved the 'displacement of concepts by quantities'.²⁴ Thus the notion of statistics, from its inception in the English language, and for at least a century after, exhibited a fluidity of meaning that lay at the core of competing claims to its importance and precision. The scope, purpose and methods of statistics were by no means fixed: to some extent this remains a foundation of controversy to this day.

Data display and collection

Important developments in the display and collection of statistics occurred in the late eighteenth and early nineteenth century. The use of graphs and visual representation date from this time although, after their first development by the Dundee businessman William Playfair in his *The Commercial and Political Atlas* of 1787, they virtually disappeared to re-emerge only slowly in this country from the later nineteenth century. Playfair's *Atlas* was well ahead of its time in including graphs of exports and imports and of English and French annual revenues and the size of the national debt, 1688–1800. His *Lineal Arithmetic* (1798) contained 37 coloured graphs covering a wider array of economic issues. *An Enquiry into the permanent causes of decline and fall of powerful and wealthy nations* (1805) soon followed with its many graphs and figures, one of which is reproduced below (Figure 2.1).

The early nineteenth century also saw the extended collection of statistical data in a variety of fields. Medical statistics were developed, largely in the work of William Black,

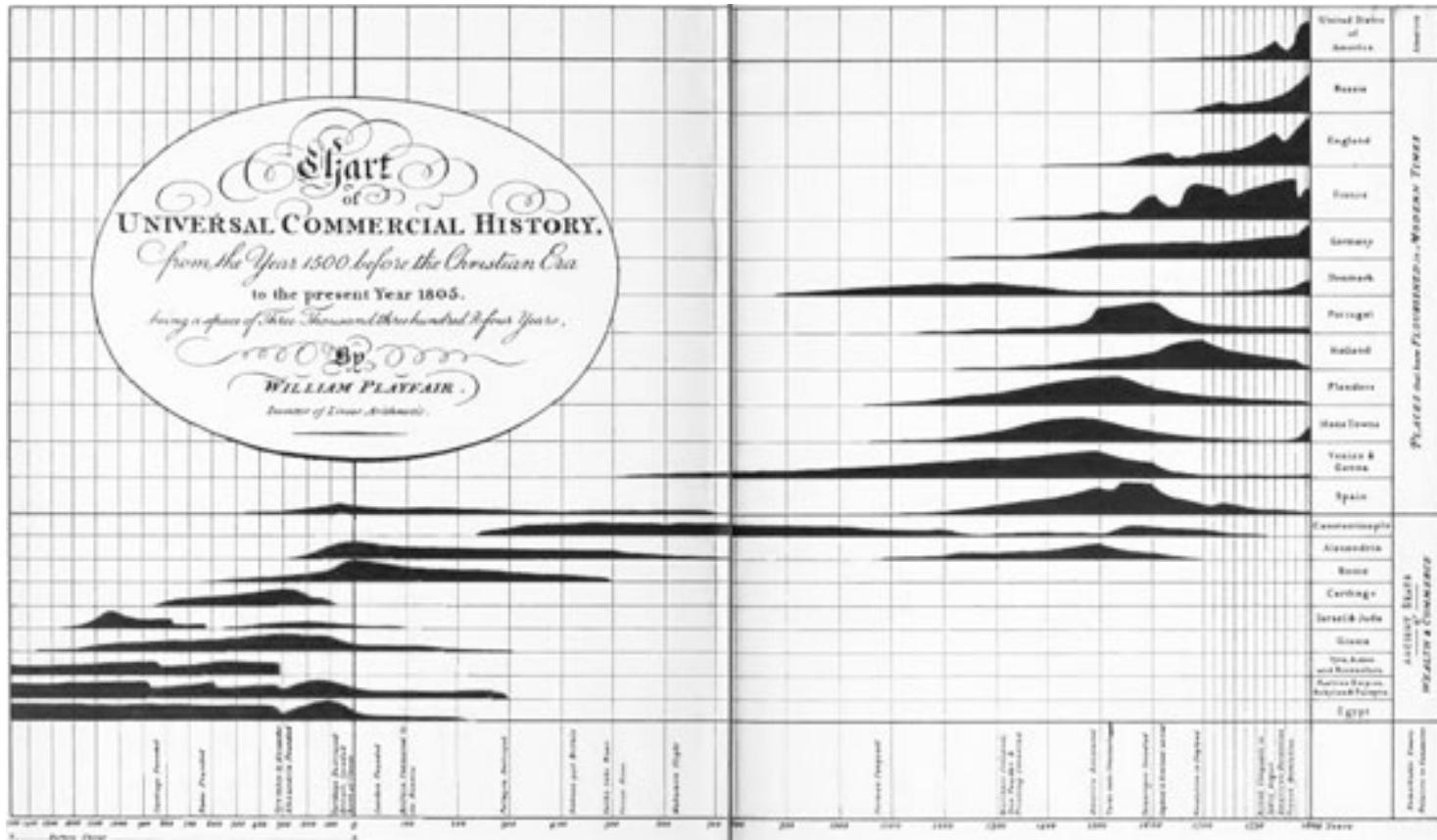


Figure 2.1 Frontispiece from William Playfair, *An enquiry into the permanent causes of decline and fall of powerful and wealthy nations* (1805).

History by Numbers

Gilbert Stone and James Annesley and in the more disparate efforts of a number of provincial medical statisticians who gathered data as a prelude for campaigning for public health reforms. National criminal statistics were published from 1810, regularly from 1832, partly because of contemporary controversy over capital punishment. Interestingly, by the time of major debates over criminal commitments which surrounded the Select Committee of 1828, it was generally agreed that most of the apparent increase in crime (which was based on the rise of commitments) arose from more effective law enforcement and from changes in the classification of crimes rather than from any real increase in offences.²⁵ Thus one of the most enduring problems of many sorts of statistics, particularly those relating to crime, was starkly exposed: the problem of distinguishing real changes from apparent shifts that arise because the numbers collected relate only indirectly to the main object of study or because the definitions or parameters of statistical measures change over time.

There were various official surveys in the early nineteenth century that generated much quantitative data and analysis: of education (1818, 1833), for example, and of factory employment (1816 Select Committee). But the most obvious and important developments in the early-nineteenth-century collection and use of social statistics were the establishment of the census (from 1801) and Civil Registration (from 1837). The census was seen as essential for effective government, for revenue raising and for recruitment of the armed forces. John Rickman, who supervised the first four censuses, had argued for the first census in order to demonstrate population increase and growing prosperity and thus to assuage the domestic discontent which characterized the inflationary and high tax years of the Napoleonic Wars. Alongside the first census, Rickman collected returns from the clergy of decennial baptism and burial figures for 1700–1780 and yearly figures thereafter. These formed a basis for some of the earliest (and much subsequent) work in historical demography.²⁶ Measures of demographic change and of the social, political and moral determinants of population growth were a key focus of numerical thinking long before Thomas Malthus published his famous *An Essay on the Principle of Population* in 1798.²⁷ But the early censuses contained only limited information derived from questions that were poorly framed. They relied on the voluntary labour of parish officials and they were inevitably highly inaccurate not least because data was collected over several days which left scope for double counting and omissions.²⁸

As in the mid-eighteenth century, it was commercial actuarial work that provided the greatest impetus to the development of statistical techniques in the early nineteenth century. This was associated particularly with Joshua Milne of the Sun Life Assurance Society.²⁹ Discontent with prevailing life tables, recognition that parish registers could not be used to rectify them because of their exclusions (particularly of dissenters) and growing demands for more accurate information about health and causes of death created major pressure for civil registration. The General Register Office was established in 1837 to administer both Civil Registration and the census and this became one of the most active government departments undertaking statistical study of social problems. By the 1840s and 1850s it was joined in the production of social and economic statistics by the Home Office, the War Office, the Board of Trade, the Admiralty and the Poor Law Commission. The statistical movement associated with the Victorian era was now well underway.

The desire for more data, and for more accurate figures to aid government and the state had parallels in France in particular. Attempts had been made to gather detailed social statistics in 1800–1811 by the *Bureau de Statistique* working through local *Intendants* who in turn relied upon local elites and savants to provide information. The result was a wonderful archive of diverse information about local landscapes, peoples, customs, festivals, dress, habits. But Napoleon Bonaparte closed the Bureau in 1811 because he needed different, more focused and uniform information for purposes of conscription, taxation and managing the war economy.³⁰ This is a good illustration of wider trends in statistical developments: a narrowing of interest to features which could more easily be categorized, quantified and compared with precision across time and space in the interests of state policy and tax gathering, but often at the expense of variety, richness and accuracy.³¹

The Victorian statistical movement

The rise of an industrialized, urbanized and commercial society was accompanied by the urge to measure, to engineer and to control. The statistical urge came partly from anxiety about social change and instability. Apart from the various offices of government busy with the production of statistical data, there developed a large number of private and voluntary associations in London such as the Statistical Society of London, the Central Society of Education, the Health of Towns Association and the Society for the Diffusion of Useful Knowledge. A plethora of reform and statistical societies and associations also sprang up in provincial towns and cities. They often promoted local initiatives as a counterweight to what was seen as central government interference and attendant high taxation. However, a widely held set of social attitudes and a common view of the needs and purposes of social reform underpinned both these associations and much of the activity of the central state. The statistical movement as a whole and the nature of data collected and classified in the nineteenth century whether by public or voluntary agencies shared a common purpose: to assist economic and social engineering and social reform.³²

Historians have been bequeathed a great deal more quantitative information for the nineteenth century than for earlier periods but there is little uniformity, standardization or sophistication in the way in which information was gathered or initially processed from the raw data. Publications of the period give a flavour of the statistical urge, the interest in trifles and oddities and the sort of ill-discipline and innovation with which statistics were collected and arranged. Figures 2.2 and 2.3 are extracts from *Mulhall's Dictionary of Statistics* (1884) which reads like a mixture of the *Guinness Book of Records* and a railway timetable. Michael George Mulhall (1829–1900) was an Irish author, economist and newspaper editor as well as a statistician. He founded the first English language newspaper in South America, as well as his *Handbook of the River Plate* (1876). Both the latter and his statistics dictionary became classics and best sellers. Arranged in alphabetical order the *Dictionary* juxtaposes statistics for apoplexy and aqueducts, hailstorms and hats, manure and marriages on the same pages and includes

F.—INCREASE IN FRANCE AND BELGIUM.

DIVORCE, PER 1000 MARRIAGES.

Period.	France.	Paris.	Belgium.	Brussels.
1826-30	1·1	4·9	1·0	4·1
1831-40	1·8	7·0	1·2	5·8
1841-50	2·8	9·9	1·4	6·6
1851-60	4·3	15·6	2·4	9·9
1861-70	5·6	22·9	2·9	11·2
1871-78	6·3	24·9	5·1	12·4

G.—DIVORCE AND SUICIDE COMPARED.

	Divorce,		Suicide,		
	per 1000 Marriages.	Inhabit.	per 1000 Marriages.	Inhabit.	
Ireland	1	17	Germany	17	14·5
England	2	67	Denmark	30	25·2
Scotland	3	4·9	Switzerland	51	20·2
Russia	2	2·5	London	4	8·6
Italy	3	8·7	Berlin	10	17·0
Sweden	7	8·1	Brussels	14	27·1
Belgium	7	7·1	Vienna	23	28·7
Holland	8	9·6	Paris	25	42·2
France	9	15·6	Stockholm	28	35·4
Austria	10	9·6	Copenhagen	29	30·2

DOCKS.—Those of London comprise 690, those of Liverpool 543, those of Cardiff 113 acres.

COST IN MILLIONS £.

London	20·1	Antwerp	6·5	Hull	1·2
Liverpool	18·2	Cherbourg	3·5	Bristol	0·9
Glasgow	7·6	Holyhead	2·0	Dundee	0·8

The new docks at Hamburg will cost 5½ millions sterling.

Dock-dues as a rule average 2 shillings a ton in European ports, the charges on a vessel of 1000 tons being as follows:—

Liverpool	£133	Hamburg	£110	Amsterdam	£81
London	125	Antwerp	93	General average	100

The largest lock in the world is that of Cardiff, 600 feet long by 80 in width, ordinary depth of water 36 feet.

DOGS.

A.—DOGS OF ALL KINDS.

	Number Lives.	Per 1000 Inhabitants.
Great Britain	1,125,000	38
Ireland	360,000	73
France	1,884,000	49
Germany	1,432,000	51

Sheep-dogs are not taxed in the United Kingdom, and the total number of dogs in the kingdom is at least 2,000,000, say 55 per 1000 inhabitants, worth £800,000. It is found that 100 male dogs go mad, as compared with 14 female. A dog accidentally locked up at Metz passed 39 days without food, and recovered.

B.—HUNTING-DOGS IN UNITED KINGDOM.

	England.	Ireland.	Scotland.	U. Kingdom.
Stag-hounds	604	246	—	850
Fox-hounds	12,866	1,522	660	15,048
Harriers	3,228	1,536	—	4,774
Beagles	448	—	74	522
Total	17,176	3,284	734	22,194

C.—BRAIN OF DOGS, IN DRAMA.

Sheep-dog	29·5	Retriever	22·7	Greyhound	22·4
Fox-hound	29·2	Collie	22·4	Terrier	20·9
Setter	26·1	Bull-dog	24·2	Spaniel	18·1
Mastiff	26·1	Newfoundland	24·9	Lap-dog	18·9

As compared with the above, the wolf has 42, the jackal 15, the fox 13 drama.

DRAINAGE.—Subsoil drainage in England costs on an average £5 per acre, and produces 5 bushels more wheat, say 20 per cent. extra. Reclaiming land in Scotland costs about £17 per acre.

For drainage of towns, see *Sewers*.

DRINK.

A.—CONSUMPTION IN UNITED KINGDOM.

Year.	MILLIONS OF GALLONS.		
	Bier.	Spirits.	Wine.
1840	640	25·7	6·5
1850	770	27·2	7·5
1871	980	33·6	16·1
1881	1,007	37·9	15·6

B.—CONSUMPTION PER INHABITANT.

Year.	Gallons.		
	Bier.	Spirits.	Wine.
1840	24·2	0·97	0·25
1860	26·5	0·93	0·26
1871	30·6	1·06	0·51
1881	28·6	1·05	0·44

Figure 2.2 Divorce, docks, dogs, drainage and drink: a page from *Mulhall's Dictionary*, 1884.

Source: Michael G. Mulhall FSS, *Mulhall's Dictionary of Statistics* (London 1884), pp. 152–153.

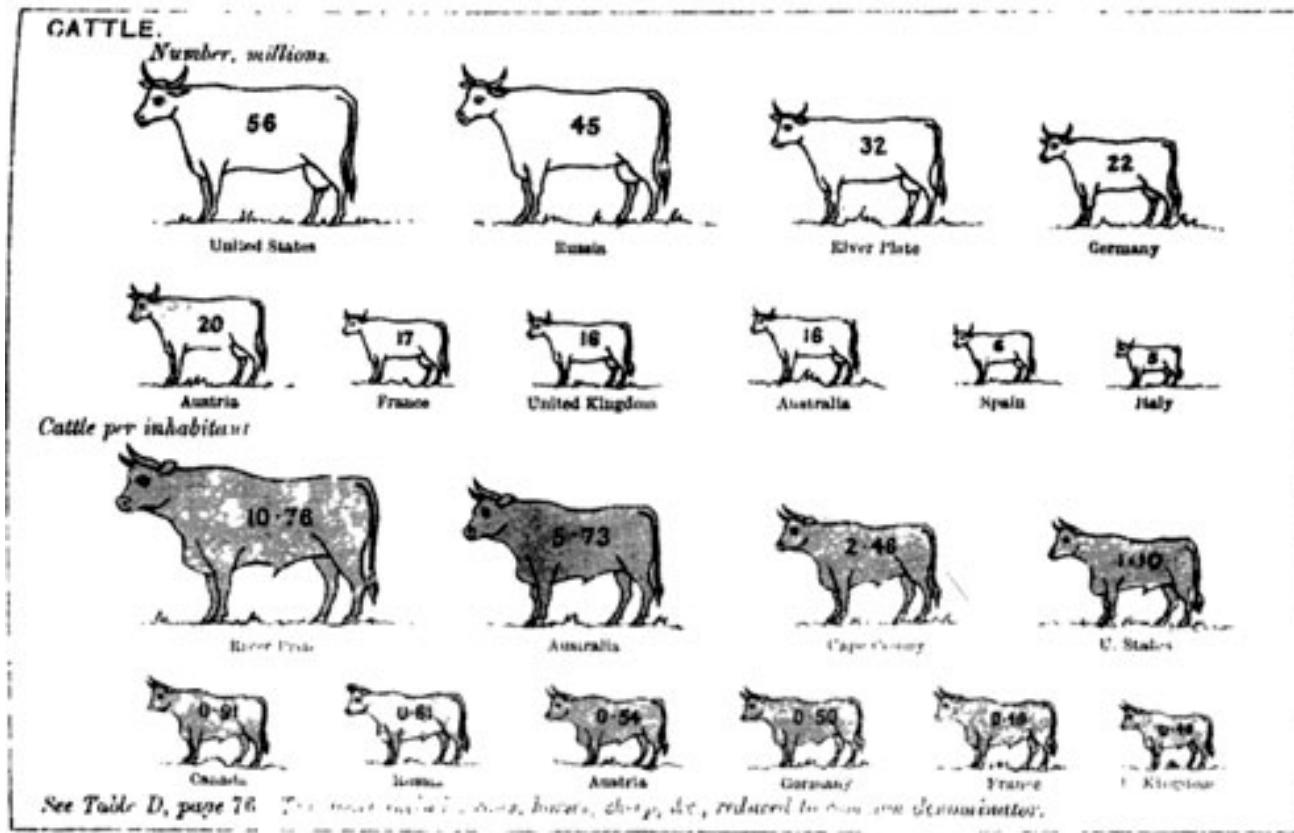


Figure 2.3 Domestic animals of the world reduced to a common denominator (cattle): pictogram from *Mulhall's Dictionary*, 1884.
Source: Michael G. Mulhall FSS, *Mulhall's Dictionary of Statistics* (London 1884), p. 77.

History by Numbers

many delightful pictograms. It can be viewed online at <https://archive.org/stream/mulhallsdiction00mulgoog#ge/n160/mode/2up>.

Much of the raw data and original evidence upon which such publications as Mulhall's were based, has not survived which means that the many printed summary figures and extracts cannot be checked. In addition, the social, political and economic preoccupations of nineteenth-century bureaucrats, civil servants and reformers conditioned and controlled the nature of the information gathered and of that left out. Even the very act of quantification itself, across the many areas of social and economic data, necessarily ignored the array of meanings and connotations that attached to things measured. As Nietzsche had pronounced: 'The form of life epitomized by quantification depends on the art of forgetting'.³³ Thus the statistical legacy that the Victorian state and voluntary associations have left for historians is a patchy one and generally difficult to use for historians' purposes. As the techniques of statistical analysis remained very primitive until the later nineteenth century, it was relatively easy for interested parties to manipulate the figures to fit particular opinions or predispositions and it is these statements (in parliamentary papers and in other contemporary accounts), backed by a *selection* of the quantitative evidence originally gathered that have most often survived.³⁴

One of the major biases of statistics gathered at this time came from the tendency amongst reformers to vindicate industrial progress by blaming social problems on other causes such as the growth of cities, alcohol consumption, the moral weakness of the poor or the evils of Anglicanism.³⁵ Statistics were often also the outcome of struggles between central and local bodies. The Board of Trade, for example, under its first director, the businessman G. R. Porter, was charged with gathering and analysing information on trade, manufactures and the economic distress of provincial England: a reaction to the commercial and social dislocation that characterized the 1830s and 1840s. But local Chambers of Commerce were unable (sometimes unwilling) to provide comprehensive or consistent information and Porter was forced to rely upon his contacts amongst merchants and manufacturers and other 'well informed gentlemen' and upon work done in a similar vein already by the London Statistical Society.³⁶ Furthermore, his brief was too disparate: it cast wide to include criminal statistics, police figures, hospital returns and there was little understanding of the value or potential for analysis and display of the figures in graphs and charts. The result was a vast mass of often impenetrable, inconsistent, partial and biased numerical information of which best use was never made in the nineteenth century and with which historians still struggle.

By 1846 the General Register Office cost £73,000 a year and employed 80 staff centrally with a further 2,800 local registrars.³⁷ This was big bureaucracy by nineteenth-century standards. William Farr was in charge of the Office and of organizing and planning the census from 1851 which was the first to include a welter of occupational and social questions. Farr's personal preoccupation with mortality and diseases, sanitary reform and improvements in occupational health drove a great deal of the statistical analysis undertaken by the Office and also determined the structure and nature of the questions framed in the census schedules and the priorities absorbed by the enumerators responsible for ensuring 'accurate' returns. Farr's work from the 1840s anticipated later developments

as it included ‘statistically controlled experiments’ designed to enable him to separate out the influences of sex, location, climate and occupation upon epidemics and mortality rates. He was also interested in the age of marriage and rate of remarriage and in the production of new life tables. Because of this the late-nineteenth-century censuses have been generally of more use and interest to demographic and medical historians than to historians interested in the family economy or the nature of work and employment of men and women as the information on these is patchy and unreliable, especially for women.³⁸

Farr represented a much wider nineteenth-century obsession with health which made itself felt in major collections of statistical information from the heights and weights of members of the armed forces, school attendees, and transported convicts. It also stimulated local and large-scale surveys of urban disease and living conditions. These have subsequently informed the history of nutrition and living standards as well as medicine, anthropometric and demographic research. At the time however, the figures were often collected and used specifically to confirm contemporary prejudices and beliefs such as the miasmic theory of disease transmission and to condemn the sorts of ‘thoughtless extravagance’ and ‘ignorance of domestic economy’ that were widely thought to lie behind the high disease and death rates of the urban working classes. Historians have to be aware and make allowances for the ways in which bias and purpose enter into the evidence that they are forced to use. Edwin Chadwick’s important surveys of urban conditions fit this pattern. Assisted in the provinces by Southwood Smith, Neil Arnott and J. P. Kay, Chadwick highlighted the ‘condition of England question’, which was seen to lie primarily in the improvidence of the poor. His analysis was a mixture of moralism and environmentalism seen in different combinations in the work of many of his contemporaries, especially those considering the health of towns. W. H. Duncan on Liverpool, Rev. J. Clay on Preston, Thomas Laycock on York and Lionel Playfair on Lancashire: all stressed the moral and spiritual as well as the environmental causes of ill health, including depraved domestic habits, improvidence and the abuse of alcohol and opiates. The term **moral statistics** was widely used in the first two-thirds of the nineteenth century highlighting the moral preoccupations of the Victorian statistical movement, preoccupations that ran through the collection and interpretation of figures on crime, education and religion as much as they did through the statistics of health.³⁹ Michelle Perrot described the French moral statisticians of the time as bourgeois reformers, seeking to control deviant behaviour of all sorts by using the power of numbers.⁴⁰ The work of Michel Foucault and his disciplinary and ‘bio-political’ models of power have been influential in alerting many historians to the ways in which statistics have contributed to the fabrication and control of society.⁴¹

Twentieth-century developments

Arising from and following upon the Victorian statistical movement, the period from the 1890s to the 1950s saw classic British statistical estimations undertaken by social commentators and historians that are still cited and used (sometimes with revisions)

History by Numbers

today. Those of Thorold Rogers, A. L. Bowley, G. H. Wood, E. H. Phelps-Brown and S. V. Hopkins, B. R. Mitchell and P. Deane are perhaps the most important.⁴² Rogers (1823–1890), a cleric and Professor of Economics and Statistics at Kings College, London was the first to attempt the collection of comprehensive long run data on the agricultural sector (six volumes from the 1860s) and on work and wages more broadly.⁴³ Increasingly, data gathering in the late nineteenth and early twentieth centuries, in many Western economies, was harnessed to the needs of political policy with predictable results. The history of the British Cost of Living Index provides an example of the way in which the purposes of data gathering influenced the evolution of major time series data that have become a pillar of historical research. Introduced in 1914 as a device to provide an evidential base for the tariff reform debate, it came to exert a major influence on many areas of government policy despite its inadequacies and despite being tied to a pre-1914 pattern of working-class expenditure for at least the next three decades. As Rebecca Searle has argued: ‘Far from being a neutral statistical measure, the official cost of living index (1914–1947) was essentially political in nature’.⁴⁴

A turning point in the statistical recording and analysis of Britain (and elsewhere in Western Europe and the US) arose from the needs of the state in the wars of the first half of the twentieth century. These necessitated a revolution in the collection of official statistics geared not, as in the past, to fiscal budgeting, but to fulfilling the demands of the war economy without inflation (by controlling domestic consumption). Growing acceptance that the state should act to stabilize the economy, to maintain economic growth and to keep unemployment low added to this mid-twentieth-century impulse for statistical record keeping, particularly at central government level.⁴⁵ The Keynesian revolution with its policy of state involvement in the economy to avoid major crises and to promote growth continued the reformist impulse already strong in the nineteenth century. The development and subsequent growth of health and social policy interventions of the Welfare State, following the Second World War further added to the demands of statistical record keeping and analysis. Statistical recording at this point remained tied largely to the idea that it would better able the government to step in and to control unemployment, incomes, economic crises, poverty and ill health, acting where the market had failed to deliver socially acceptable outcomes.

Since the mid-twentieth century new and rapidly growing sources of statistical evidence about society and the economy have impacted upon the use of quantitative evidence in historical research on the modern period and, in turn, have influenced historical research methods. It is no accident, for example, that the composition and growth of GNP in different phases of economic development became a central concern of economic history from this time relating both to the newly available source materials and to developing methods of contemporary national accounting. Improved censuses of production influenced economic history. The evidence arising from new social data such as consumer expenditure surveys revolutionized the ways in which social historians were able to conceive and to analyse changing tastes and fashions whilst new data on mortality and diseases had a similar effect upon the history of medicine. Whilst state interventionist motivations in statistical development remained strong the neo-liberal political turn in Western countries in the later

twentieth century saw statistical data and techniques increasingly devoted to research, both contemporary and historical, that was geared increasingly to analysing and measuring the efficiency of markets and to calibrating the impact of innovations and institutions that either promoted or inhibited entrepreneurship and free trade.

Desrosieres has argued that, from the 1930s there was a ‘co-construction of techniques of governing, data accumulation and the logics of abstraction’.⁴⁶ In other words, data collection and statistical practices, financed by the state and used for or geared to state purposes not only multiplied in the mid-twentieth century but also, in turn, influenced statistical theory to which we now turn.

Statistical theory

The later nineteenth and twentieth centuries saw not just the continued collection of masses of socially constructed and socially interpreted quantitative data but also witnessed major developments in statistical theory and analysis that, with relatively minor refinements, are still the central elements in statistics to this day. Understanding how and why these new techniques (and those that followed) arose is interesting in itself but it is also important in illustrating that statistical and other ‘scientific’ procedures cannot be seen as separate from the environment in which they were formed and the purposes that they were geared to serve. They were socially as well as scientifically constructed.⁴⁷

By the mid-nineteenth century there was growing recognition of the regularities that were emerging in social data. In his opening address to the Royal Statistical Society, published in 1860, Nassau Senior (who had been instrumental in investigating and transforming the Poor Law in 1834) suggested that ‘the human will obeys laws nearly as certain as those which regulate matter’. The most influential thinker to pursue a numerical social *science* that stressed the regularities and periodicities to be found in social data, and the use of probability theory in understanding these, was the astronomer Adolphe Quetelet. He drew parallels between statistics and astronomy and studied aggregated figures of phenomena such as mortality, crime and suicide. He saw statistical regularity as providing the key to social science and promoted the ‘law of large numbers’ which implied that general effects in society are always produced by general causes, accidental or chance factors having no influence when a mass are considered collectively. His most celebrated construct was *L'homme moyen* or average man, an abstract person who exhibited all the human attributes in a given country. Quetelet’s belief that statistical laws could prevail for a mass even when the constituent individuals were too numerous or too inscrutable for their actions to be understood individually, became the foundation of much later social science theory, especially in economics, and was also important in influencing ideas in the natural sciences.⁴⁸

Marx used Quetelet’s notion of the average man in defining the labour theory of value and, most notably, Emile Durkheim used the statistical ideas of Quetelet and others in his research on suicide. In *Suicide* (1897), the first great sociological work based upon quantification, Durkheim used statistics to show that people were driven by collective

History by Numbers

impulses which could not be reduced to the particular circumstances that individuals cite to rationalize their deeds.⁴⁹ In history the idea that large datasets could provide insights into the functioning of society was taken up in the work of Henry Thomas Buckle who denounced the mediocre presentation of chronicles of kings and battles and aimed to inject some science, order, symmetry and law into history.⁵⁰

During the later nineteenth century confidence in the value and reliability of statistical laws spread from the social to the physical and biological sciences and the analogies and similes of social science were used frequently in science from thermodynamics to heredity. Approaches in these disciplines were united by the idea that the greatest variability and chaos at one level was consistent with remarkable stability at the aggregate level, which manifested itself in the statistical laws of large numbers.

The development of statistical procedures for dealing with aggregated phenomena came to dominate the field for a century in both natural and social sciences. It was associated with the work of three men in particular: Francis Galton (1822–1911), Karl Pearson (1857–1936) and R. A. Fisher (1890–1962). Galton was a cousin of Charles Darwin, and invented correlation and regression analysis, providing the major breakthrough which enabled simultaneous analysis of the movement of more than one variable. Pearson, a Fabian socialist and atheist, developed and systematized Galton's insights, gave his name to the product moment coefficient of correlation, produced the chi-squared distribution and founded the biometric school.⁵¹ His son Egon Pearson continued this role as he also continued the friction that had developed over method between his father and Ronald Fisher. Fisher reshaped the basis of statistical theorizing by systematizing the analysis of variance and by creating tests to indicate the significance of research results based on evidence from discrete experiments or samples (see below). He was actively involved in agricultural and biological experiments.⁵²

All three, despite their different orientations, were eugenicists. They were from a much wider body of influential scientists and politicians and other professionals of the time who claimed that the most important biological characteristics, including human traits such as mental ability, were all inherited and that ancestry, rather than environment or mutation, was the crucial variable in determining both intelligence and behaviour. It is possible that moral statistics was an influence in this line of argument but the conclusions drawn went beyond the moral condemnations of Chadwick and others to suggest that the only long-term way to solve the problems of an expanding commercial society was to improve the bloodstock to ensure that those with good characteristics (the fit) had more children than those with bad (the unfit). At the root of this was the view that the urban poor were morally degenerate as well as dangerous. Natural selection had failed to root them out because of the intervention of charity, medicine and sanitary reform. This body of ideas became notorious and controversial in the twentieth century in large part, but not solely, because of the racial and ethnic policies of the Nazi Party in Germany which relied heavily upon eugenicist ideas. Debates about race, class and IQ continue to spark vitriolic debate especially where supported by quantitative data.⁵³

The racial and class assumptions of eugenics were deeply embedded in late nineteenth-century social and scientific ideas, particularly amongst the professional middle class, in

law, medicine and academic life. Karl Pearson, for example, was elected to the new Chair in Eugenics at University College, London in 1911. The manipulation of numbers in seemingly objective scientific enquiry assisted in placing eugenics in a respectable light and indicates how dangerous numbers can be unless their use is policed and questioned by people unafraid to get involved in disputes over seemingly incontrovertible, ‘scientific’ evidence. Galton and Pearson developed statistical theory to apply to their eugenics research for studies of twins, correlations of intelligence between relatives, measures of hereditability and racial inferiority. It has been argued, particularly by Donald A. Mackenzie, that eugenics did not merely motivate their statistical work but affected both its content and its methods. The shape of the science they developed was partially determined by eugenic objectives and these objectives also determined new ways of collecting and storing data.⁵⁴ Because their central concern was with the impact of the characteristics of one generation on the next, the statistical dependence of pairs of variables (height, intelligence or whatever) were central to their research. Pearson’s work on statistical association clashed with that of his former pupil George Udny Yule. Pearson’s need for the *measurability* of associations between variables that might be inherited differed from Yule’s desire merely to indicate whether there was any *indication* of an association or not.⁵⁵

There was also a bitter personal controversy between Pearson and R. A. Fisher. Unlike Pearson, the much more conservative Fisher believed that hereditability of characteristics was entirely a result of genetics and that this vindicated the rigour of an entirely statistical, probability-based approach to the subject, termed ‘biometric eugenics’.⁵⁶ Arguably the most important element of statistical theory developed by Fisher was in relation to statistical inference: the ability to infer from the known and the examined to the unknown and the unexamined. The most widely employed idea concerning statistical inference before Fisher had been Bayes’ theorem, known as the method of inverse probability. Bayes’ theorem had suggested that we should change our belief in a theory in the light of new evidence and its effects on the probability that an initial idea was correct.⁵⁷ The main problem for Fisher and others was that prior probabilities and existing beliefs in relationships or theories varied from person to person and were hard to pin down. This struck at the heart of the scientific community’s fear of subjectivity. Fisher claimed to provide an objective alternative. In his view, the accuracy of a scientific finding should be judged in relation to the probability of getting results at least as impressive as those obtained assuming they had occurred entirely by chance. A probability value P of 0.05, that is, one in 20, was arbitrarily accepted as reasonable and this probability has assumed enormous importance in statistical inference ever since even though it has been argued that the technique ‘routinely exaggerates the size and significance of implausible findings’.⁵⁸ Where Bayes’ theorem took account of context, prior knowledge and plausibility, Fisher’s test has no means of taking subjective elements including plausibility into account. Yet it is still Fisher’s test that is dominant today in this field.⁵⁹ It has been modified by researchers to include confidence intervals (that is, likely ranges of error in the results) but this still does not factor in plausibility. This underlines the fact that statistical tests should be undertaken when and only when there is a highly plausible

History by Numbers

reason for doing so (based on experience and prior knowledge) and that statistical significance should *never* be equated with substantive or real significance.⁶⁰

The drive to quantify and record so many aspects of economic and social life in the nineteenth and twentieth centuries – to know and thus to control society as well as the natural and physical world – was reflected in the nature of both economics and sociology as they emerged as academic subjects in the late nineteenth and early twentieth century. It predisposed them, from the outset, to use quantitative methods. In history, however, much less emphasis was placed upon statistical and mathematical analysis, outside of economic history, until well into the twentieth century. This is because quantification, from its inception, was closely associated with a particular understanding about knowledge and its construction: positivism. This suited social science, economics and social policy orientations far better than history.

Positivism

Positivism assumes that the only true knowledge is scientific knowledge that describes and explains observable phenomena. The assumption behind positivism is the possibility of neutral and meticulous observation of facts that eventually reveal regularities and even laws of behaviour. In the nineteenth century the dominant belief within the scientific and intellectual community was that close observation of the recorded facts, would lead, by a process of induction, to an understanding of 'laws' about the workings of the economy, society and historical development just the same as in the natural sciences. The approach of Darwin's *The Origin of Species* could be followed in the study of human social evolution. The idea was that scientists of society like those of physics or chemistry, approached their task without preconceptions or moral involvement, gathered evidence neutrally, applied rules and criteria to interpret the evidence, and drew their conclusions from this. Through study and induction, positive knowledge of social phenomena would thus provide the basis for scientifically grounded intervention in economic and social affairs to the benefit of society as a whole. The positivist desire for quantification was powerful because of the widely shared social ambitions for science.

Nowadays scientific knowledge is seen less as the build-up of some all-encompassing body of realist and objective knowledge than of the rise and decline of successive ways of theorizing about the physical world (the rise and decline of paradigms of thought). It is also regarded as the study of probabilities rather than of laws or absolute truths. This is partly because of the influence of statistical approaches in both natural and social science. The meaning of objectivity has shifted at the same time. It is now uncommon to believe that objectivity can be identified with 'truth' or knowledge congruent with the real. Instead, objectivity is applied to knowledge that meets criteria of validity and reliability that are held to be as free from bias as possible. It indicates the avoidance of subjectivity by following 'impartial' rules of measurement, observation and experiment. This is in no sense unrelated to the disciplines of quantification and statistics. Karl Pearson was the type of positivist who argued that there is no knowable thing in itself

underlying our perceptions and that perceptions provide the whole basis of knowledge. But for him and others this made the rigour and objectivity of testing perceptions for their accuracy all the more important. Statistics and statistical theory gained ground and became a prominent discourse in many fields of enquiry not despite the shift from belief in underlying truths but because of it.

The notion that there is an internal logic of scientific development unaffected by its social context may have been popular in the nineteenth century, but it is now regarded with great scepticism. Similarly in history there is now greater recognition of historical relativism: the acceptance that historians are deeply influenced by their own culture and environment and that this fundamentally conditions the history which they write. Thus, it is generally acknowledged that all history is to a greater or lesser extent contemporary history in that it reflects the views and interests of contemporary writers as much as, if not more than, it tells us about the past: history is researched and written through the eyes, the preoccupations and the language of later periods. Despite or because of these changes in understanding, there is still a great deal of friction amongst historians and social scientists about the degree to which they are involved in a scientific methodology. And this friction has increased in recent years in those areas of the humanities influenced by post-modernism, post-structuralism and an emphasis upon language and discourse as the beginning and the end point of knowledge. Some social scientists and historians now claim that we have left behind the old modernist certainties of building up a knowledge of social life, past or present, from our attempted observations of ‘reality’. Instead they emphasize the role of ideas and discourse in creating knowledge and deny any direct relationship between knowledge and reality.⁶¹

The association of quantitative approaches with positivistic scientific enquiry and with attempts to mould society along particular lines, made them the brunt of post-modern condemnation in the closing decades of the twentieth century. For post-modern critics, numbers represent the cutting edge of modernity. Numbers were, and often are still, seen as acting on people and exercising power over them particularly through the creation of statistically verifiable behavioural norms against which an oppressive category of abnormality is created.⁶² Yet this is perhaps unfair because quantification shares with other ‘languages’ many of the same characteristics in helping to create the society it purports to describe or to mediate. Other languages also share, with quantification, the production of oppressive categories that promote or extol individuals or behaviours by condemning their opposites. The fact that quantification appears, and is promoted as, more objective and more rigorous in its hold on reality than qualitative accounts, and that it is potentially more hegemonic across time and space than other languages, are dangers and call for vigilance. But these should not blind us to problems that occur with other approaches to history. The force of linguistic structures, manners of delivery, rhetoric, style and narrative need similar evaluation, care and reflexivity. One could argue that words carry subtleties of connotation and context and fertile ambiguities which allow one to communicate more effectively than in numbers. But what words gain in flexibility they lose vis-à-vis numbers in precision. Both carry problems and a mix of the two appears desirable. Unfortunately, we are too often presented with a mutually

History by Numbers

exclusive choice between words and numbers (qualitative and quantitative approaches) as if each represents an entirely different form of knowledge. As Anthony Giddens has argued ‘all quantitative data, when scrutinised, turn out to be composites of “qualitative” ... interpretations, produced by situated researchers, coders, government officials and others.’⁶³ The problems that quantification shares with other discourses is one of degree rather than kind. ‘The credibility of numbers, like the credibility of knowledge in any form, is a social and moral problem.’⁶⁴

Objectivity and prejudice

Objectivity, and the desirability of objectivity, is a key to understanding what numbers and their disciplined application have in common with words within a linguistic structure. Objectivity is sometimes understood to refer to accounts of the external world that are held to represent the world as it exists independently of our conceptions. But a much more frequent usage is in reference to knowledge claimed to meet criteria of validity and reliability and held to be as free from bias as possible. As few intellectuals these days would subscribe to the view that objectivity means truth or an unclouded appreciation of reality, the second meaning (**mechanical objectivity**) is the one which should concern us here and it is largely its use in this sense that has driven quantifiers in their endeavours and that has helped to secure the authority both of science and of numbers. In this sense objectivity means personal restraint and following rules. This encapsulates the way in which statistical methodology has evolved, particularly amongst its positivist practitioners and pioneers in the late nineteenth and early twentieth centuries. Pearson, for example, was wedded to the subordination of personal interests and prejudices to public standards, and to the moral training and self-denial that this involved. He believed that mathematical statistics should be the language of reasoning in all areas of human activity especially in government which had for so long been in the hands of scientifically illiterate aristocrats. Pearson’s goal, and one that he saw as more important than any idea of objectivity as truth, was the spread to all aspects of life of an ordered method of investigation and the taming of individual subjectivity and bias in the interests of society. This attitude can be seen pervading the growth of quantification in the nineteenth and twentieth centuries. Often the accuracy of observations, which might employ experience and intuition, has been subordinated to the need for a common set of categories of analysis, a common language and comparability of results.

This background helps one to understand why there has been, and still is, so much bad feeling between the more rigorous exponents of quantitative history, and those who condemn much of the work done in this area. The clash over quantification can be seen as philosophical as well as practical: covering the nature of history as well as of appropriate research techniques. But antipathies are also fuelled by mutual misunderstandings, ignorance and prejudice. Quantitative history in itself need be no less nor no more positivistic or scientific than an analysis based on qualitative data. Indeed the two approaches are by no means as methodologically separate or distinctive as much debate

has made them out to be. In opposing exchanges between those who support and those who deride quantification in history it is frequently the case that each side creates a caricature of the other. In particular those opposing quantification too often imply that all quantitative history (where it is not pure mystification) is an attempt to impose an inappropriate scientific methodology upon the analysis of complex human behaviour in the past: to force evidence into classificatory straightjackets which allow too little for diversity or unpredictability. Alternatively, it is claimed that the statistical techniques themselves are so imbued with the values and prejudices of those who were responsible for creating them that they are of little use in wider contexts. But it would be a mistake to think that theories and techniques of statistics are any more tainted in this way than other sorts of theories, ideas and concepts used in social study. In fact some recent thawing of the qualitative/quantitative conflict in the social sciences owes much to discussion of the reflexive, subjective and normative nature of qualitative approaches and of the problems inherent in language and linguistic structures which circumscribe and define what is said, written and thought. Whether we communicate with each other in words or in numbers or in a mixture of the two, many of the problems are the same. The so-called 'linguistic turn' in recent decades that involved historians looking closely at the values, beliefs, assumptions and categories embedded in the language of documents and the language used by historians, has a counterpart in an older critique of quantification and its power to restrict understanding by narrowing the scope of discourse to things neatly and subjectively categorized and enumerated.

Many statistical applications are a useful aid to the display or summary of facts relevant to an argument. They are often used alongside other sorts of approaches using qualitative evidence and are no more scientific or positivistic for being numerical than would be a descriptive section of prose used in the same context. At the opposite extreme the application of a model of human behaviour to the past and the empirical testing of such a model would constitute a 'scientific' methodology whether or not the model and its testing were framed in quantitative or rhetorical terms. More often than not social science models, especially those derived from economics, do involve numerical analysis. This accounts for the common but often mistaken identification of positivistic scientific approaches with quantification. But the debate about whether history can be regarded as a (social) science is not the same as that concerning the advantages or otherwise of quantitative and non-quantitative history. Although some quantitative historians make extreme claims to greater objectivity and analytical rigour than can be possible with other techniques, this is not necessarily a hallmark of quantification in itself.

Conclusion

From this chapter we have learned about the rise of data collection, quantitative thinking and analysis and we have considered the close connection between these and the goals and assumptions of their creators and the worlds in which they lived. We have learned that quantification is not merely a strategy for describing the social (and natural) worlds but a

History by Numbers

means of reconfiguring them. But qualitative approaches are not intrinsically superior in this respect because, like quantification, they also use socially constructed categories and rules of communication. Qualitative approaches may be less precise and more multivalent than quantitative approaches but neither has an intrinsically superior claim to distance from the predispositions and beliefs of their creators despite protestations to the contrary.

There are no easy answers these days to the question what is history? But, in their efforts to understand the past, historians are not helped by a polarization of opinion about quantitative and non-quantitative methods. There is a wide spectrum of quantitative evidence, and many useful, often simple, techniques that can be used in historical research, providing one remains aware of the pitfalls and biases of the evidence. There is a similar spectrum of sources, concepts, theories, methods and pitfalls which underpin qualitative history. Each piece of research, whether relying heavily on numbers or not must be judged on its own merits: by the consistency and cogency of arguments in relation to a critical use of the available evidence. A critical approach to the social construction of evidence and 'knowledge' and a reflexive attitude on the part of the historian are essential whether we are considering quantitative or qualitative history. But this involves leaving behind what has become a rather sterile and unhelpful debate about the *inherent* superiority of one approach over the other.

Further reading

- Cullen, M. J., *The Statistical Movement in Early Victorian Britain* (New York 1975).
- Daston, L., *Classical Probability in the Enlightenment* (Princeton, NJ 1988).
- Deane, P., 'Political arithmetic', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: A Dictionary of Economics* (4 volumes, London 1987), pp. 990–993.
- Desrosieres, Alain, *The Politics of Large Numbers. A History of Statistical Reasoning* (Cambridge, MA 1998).
- Desrosieres, Alain, 'Managing the economy', in T. M. Porter and D. Ross (eds), *The Modern Social Sciences* (Cambridge, MA 2003), pp. 553–564.
- Frängsmyr, T., J. L. Heilbron and R. E. Rider (eds), *The Quantifying Spirit in the Eighteenth Century* (Oxford and Los Angeles 1990).
- Glass, D. V., *Numbering the People: The Great Demography Controversy* (London 1978).
- Hacking, I., *The Taming of Chance* (Cambridge 1990).
- Higgs, E., *Making Sense of the Census* (London 1989).
- Higgs, E., *The Information State in England: The Central Collection of Information on Citizens Since 1500* (Basingstoke 2004).
- Hoppit, J., 'Reforming Britain's weights and measures, 1660–1824', *English Historical Review*, 108 (426), (1993), pp. 82–104.
- Hoppit, J., 'Political arithmetic in eighteenth-century England', *Economic History Review*, 49 (3), (1996), pp. 516–540.
- Hudson, Pat, 'Numbers and words: quantitative methods for scholars of texts', in G. Griffin (ed.), *Research Methods for English Studies* (Edinburgh 2005), pp. 131–156.
- Kula, W., *Measures and Men* (Princeton, NJ 1986).
- Mackenzie, D. A., *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge* (Edinburgh 1981).
- Matthews, R., 'Flukes and flaws', *Prospect*, November 1998, pp. 20–24.

The Origins and Nature of Quantitative Thinking

- Norton Wise, M. (ed.), *The Values of Precision* (Princeton, NJ 1995).
- Patriarca, Silvana, *Numbers and Nationhood: Writing Statistics in Nineteenth-Century Italy* (Cambridge 1996).
- Pearson, K., *The history of statistics in the seventeenth and eighteenth centuries against the changing background of intellectual, scientific and religious thought*, ed. E. S. Pearson (London 1936–1938; reissued Lubrecht & Cramer Ltd, London 1978).
- Porter, T. M., *The Rise of Statistical Thinking, 1820–1900* (Princeton, NJ 1986).
- Porter, T. M., *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton, NJ 1995).
- Rusnock, Andrea A., *Vital Accounts. Quantifying Health and Population in Eighteenth-Century England and France* (Cambridge 2002).
- Schweber, Libby, *Disciplining Statistics. Demography and Vital Statistics in France and England, 1830–1885* (Durham, NJ 2006).
- Stigler, S. M., *The History of Statistics: The Measurement of Uncertainty Before 1900* (Cambridge, MA 1986).
- Stone, J. R. N., *Some British Empiricists in the Social Sciences, 1650–1900* (Cambridge 1997).

CHAPTER 3

ARRANGING, REARRANGING AND DISPLAYING DATA

The purpose of Chapters 3 and 4 is to introduce the historian to the nature of different sorts of raw quantitative evidence and to ways of arranging and studying that evidence so that important questions can immediately be asked and answered. This aspect of quantitative method is generally referred to as **elementary descriptive statistics**. It is concerned with identifying the most important features of the data through rearrangement and presentation.

After gathering data relevant to a research project, the next step will usually involve a close look at what the evidence might initially reveal. This is assisted by rearranging the material into an easily understandable and accessible format geared to the questions posed. Much can be learned simply from the arrangement and display of data into clearly labelled tables and figures of various kinds, providing such arrangements and displays are done with care to avoid introducing distortions. Most historical research published in journals and monographs today uses quantitative evidence, if at all, in this manner, hence the importance of mastering the simple procedures outlined here.

Types of data

It is important at the outset to classify data by type because quantitative analysis can be undertaken using some sorts of data and not others. There are two main types of data: **categorical** and **numeric** and it is common to recognize two subclasses in each of these, as highlighted in Table 3.1.

Table 3.1 Types of variables

Variable type	Description	Examples
Categorical		
Nominal (unordered)	Gives only qualitative information	Names, occupations, nationalities, sex, religion
Ordinal (ordered)	Ranking or order is important	Social status, economic class
Numeric		
Interval	Distance between values has meaning	Year, temperature
Ratio	Ratio of two values has meaning	Wealth, age, prices, wages

Source: Based on Loren Haskins and Kirk Jeffrey, *Understanding Quantitative History* (Cambridge, MA 1990), p. 211, Table 6.1.

History by Numbers

Nominal data

The term ‘nominal’ comes from the word name. Common nominal variables are names of persons, places, institutions (like school or university), possessions, nationalities, religions, commonly occurring words in a text. Nominal data are always discrete, that is separate, distinct and individual. Some variables that can only take two ‘values’ (so-called dichotomous variables), such as whether male or female, or whether married or unmarried, may also be regarded as nominal. The defining character of nominal data (sometimes also termed categorical data) is that the variable cannot be rank ordered. In other words, rearranging the order of the listing of a nominal variable would have no point and would not affect the level of information provided. Also it is not possible to add up a list of nominal data as the units of expression are varied. Lists of individuals, business enterprises or occupations are common examples of nominal data. However, if they are ranked in any definable order by wealth, importance or size they should more properly be regarded as **ordinal data** (see below).

Sometimes nominal data are coded to make processing easier. For example nationalities might be coded as Swedish 1, Canadian 2, French 3, German 4 and so on with no significance attached to the numerical order or magnitude: it is just a reference code. Where this occurs the data remain nominal (or categorical) even though they are expressed numerically. This can cause some confusion but a quick rule of thumb is to ask ‘would it make any sense at all to take an average of these numbers?’ If the answer is no then the series of numbers are merely codes representing nominal or categorical variables.

Table 3.2 shows two sorts of nominal variables in relation to one another: the gender of victims of theft in 1800 and the sort of crime to which they were subject. The data is drawn from one of the most-used electronic datasets currently available relating to the social history of London: Old Bailey Online (www.oldbaileyonline.org). Old Bailey Online contains machine searchable records of all court cases at the Old Bailey 1674–1913, including the full transcripts of 197,745 criminal trials. In Table 3.2 below based on a very small subset of cases in the year 1800, in neither case is the order in which the nominal variables are placed of any significance.

Table 3.3 again illustrates two sorts of nominal variables in relationship to one another: income group and the objects of their household expenditure in Britain in the 1930s. The relative importance of expenditure on different items varies with social group

Table 3.2 Numbers of victims by gender of different thefts committed by female defendants in 1800

Victim gender	Larceny	Theft from a specified place	Shoplifting	Pocket picking	House breaking	Burglary	Receiving
Male	63	18	11	8	1	1	
Female	15	4					
Unknown	14	3	2		4		16

Source: Old Bailey Proceedings (www.oldbaileyonline.org).

Table 3.3 Average weekly expenditure for non-agricultural and agricultural working-class households, and for middle-class families with a head of household earning £250–350 per year, 1937–1938

Mean weekly expenditure for households in the:	Low income middle class		Non-agricultural working class		Agricultural working class	
	Shillings	%	Shillings	%	Shillings	%
Food	35.6	26.2	34.1	40.1	27.8	48.4
Accommodation	19.0	14.0	10.8	12.7	4.8	8.3
Clothing & footwear	12.4	9.1	8.1	9.5	5.3	9.2
Fuel & light	8.5	6.2	6.4	7.5	4.9	8.6
Other items:	60.5	44.5	25.6	30.1	14.6	25.5
Household items	10.9	8.0	4.1	4.8	2.3	4.1
Tobacco & cigarettes	2.8	2.0	2.5	3.0	1.9	3.3
Transport	4.1	3.0	2.3	2.6	0.9	1.5
Medical, insurance, pensions, union subscriptions	15.4	11.3	7.5	8.8	4.8	8.4
Other	27.5	20.2	9.3	10.9	4.7	8.2
Total weekly expenditure	136.0	100.0	85.0	100.0	57.3	100.0

Source: The National Archives LAB 17/7, July 1949. Based on Francesca Carnevali and Julie Marie Strange (eds) (1994) *Twentieth Century Britain: Economic Cultural and Social Change*, Harlow: Pearson Education, p. 164.

so the order in which they appear in the list on the left hand side of the table is not important (it involves no hierarchy). Income groups at the top of the table are not ordered by social or income status so this is again nominal rather than ordinal data. Had the different elements of the working and middle income classes been ordered in a hierarchy in terms of total weekly expenditure, that is, had ‘agricultural working class’ (with lower total expenditure) preceded ‘non-agricultural working class’ before moving on to ‘middle class’, the data might have been considered to be ordinal.

Ordinal data

The term ordinal comes from the word ‘order’. Ordinal data are generally more informative than nominal because the order is important. The order indicates relative size, status, age or some other hierarchical feature. In other words, *ordinal data represents a hierarchy of information*. For example:

- Listings of occupations in order of social status
- Wealth holding groups by value of assets
- Wage bands or earnings rankings.

An example of ordinal data is provided in Table 3.4 which represents a much used ranking of income (and potential wealth) of elite families in 1688, produced by Gregory

History by Numbers

King. Unlike nominal data, if the order here was disturbed an important additional piece of information would be lost: the social status hierarchy.

Another example of ordinal data is given in Table 3.5 which gives information on different textiles in the wardrobes of various social groups in eighteenth-century Paris, at two different points in time. The various textiles are nominal data because their order is irrelevant. Because the different social classes are arranged in a hierarchy from left to right in the columns of the table, the social class ascriptions are in this case ordinal.

With ordinal data, *order* is important but not the difference between the values: it is impossible to state with certainty whether the intervals between each value are equal or have any meaning.

Table 3.4 Numbers in social classes, c. 1688

Class	Number of families
Temporal lords	160
Spiritual lords	26
Baronets	800
Knights	600
Esquires	3,000
Gentlemen	12,000
Persons in greater offices and places	5,000
Persons in lesser offices and places	5,000
Eminent merchants and traders by sea	2,000
Lesser merchants and traders by sea	8,000
Persons in the law	10,000
Eminent clergymen	2,000
Lesser clergymen	8,000

Source: Gregory King, quoted in L. Soltow, 'Long run changes in British income inequality', *Economic History Review*, XXI, 1 (1968), p. 18.

Table 3.5 Fabrics in wardrobes of Parisians of different classes in 1700 and 1789 (per cent)

Fabric	Nobles		Professionals		Artisans & shopkeepers		Wage earners	
	1700	1789	1700	1789	1700	1789	1700	1789
Silk	17	38	17	31	13	21	9	15
Wool	8	18	22	23	23	23	58	33
Linen	46	17	37	13	42	12	14	12
Cotton	7	25	3	20	8	39	7	38
Miscellaneous	22	2	21	13	14	5	12	2

Source: Based on William H. Sewell Jr (2010) 'The empire of fashion and the rise of capitalism in eighteenth-century France', *Past and Present*, 206 (1), 109.

Interval data

Interval data are always numeric and the distances or intervals between the numbers often have meaning. Examples of interval data include dates, heights, weights, wages, attendees, numbers of persons or items in different categories.

Interval data is either counted in whole numbers as with year dates or numbers of people (**discrete data**) or measured on a continuous scale (**continuous data**) as with heights, weights, wages or dates (that are not rounded to the nearest year or decade, for example).

Table 3.6 illustrates discrete interval data (by hierarchical room value range) in the percentage of various coloured bed hangings in London in the later seventeenth century.

Table 3.6 Colour analysis from bed hangings, 1660–1675, by room value range and room name (per cent)

Colour	Room Value Range					From Room Name
	Under £10	£10–20	£20–30	Over £30	Combined	
Green	46	38	25	20	38	37
Red	23	26	32	20	25	29
Blue	16	7	9	3	11	11
Purple	4	13	5	23	9	10
Sad colour	4	8	13	17	7	3
Grey	3	4	5	3	4	5
Yellow	2	1	4	3	1	1
White	2	1	5	0	1	3
Others	0	2	2	11	4	1
Total	100	100	100	100	100	100

Source: Based on D. M. Mitchell (2009) ‘My purple will be too sad for that melancholy room: furnishings for interiors in London and Paris, 1660–1735’, *Textile History*, 40 (2), 16.

Ratio data

Ratio data is interval data where the ratio of values has meaning: for example, with ages or incomes. Ratio variables always have a zero value on their scale. Most data used in quantitative analysis are of ratio type because the additional information (common units of measure and known equal intervals between classes), as well as the existence of a zero value, is a necessary prerequisite for most further statistical procedures.

Examples of ratio data include:

- yields of crops in bushels or some other common unit of measure;
- wages or earnings data in £ s. d.;
- price movements over time for particular goods or services;

History by Numbers

- series of imports or exports by value over time;
- prison sentences for different crimes or from different courts;
- numbers of children born each year;
- numbers of times a particular word or phrase appears in a discrete range of texts.

Sometimes it is hard at first to spot the type of data. Nominal data if arranged in a meaningful order become ordinal data. Similarly, the dividing line between ordinal and ratio data is sometimes confusing: the guiding point should be that *in ratio data there must be common units of measure which can be added, divided and averaged.*

In Table 3.3 above we mentioned that two sets of nominal variables are arranged in relation to one another but additionally the level of expenditure of different social groups on different items is given in the common unit shillings. The shillings columns show ratio data and the percentage columns demonstrate the usefulness of being able to add the sums and divide by 100 to give a percentage for each reading.

Some definitions involved in regrouping data

Once the nature of the data has been identified, reclassifying and regrouping into a data matrix is a common first step that the historian takes in order to display the information more effectively and before undertaking any sort of statistical *analysis*. Some definitions are important here:

A **dataset** is a group of data selected or gathered by the historian to help him or her answer a particular question.

The dataset usually consists of a series of **cases**.

A case consists of one or more pieces of information relating to a particular unit of investigation. For example, if one had a list of women with their names, ages and occupations, each distinct woman, with the various pieces of information pertaining to her, would constitute a case. Similarly if we had a list of novels with author name, title, language, word count, then each novel, together with the associated information, would constitute a case.

Each piece of information relating to a case is called a **variable**. In other words a variable is a characteristic or a measure associated with each case in the dataset. The characteristic or measure is likely to vary from one case to the next hence the term variable. In our list of women with their ages and occupations the variables are names (nominal), ages (ratio) and occupations (nominal provided they are not ranked in order of prestige or income in which case they would be ordinal). In our list of novels, the variables are author (nominal), title (nominal), language (nominal), word count (interval/ratio).

A **data matrix** is a tabular representation of the data. It is a convenient way of organizing and tabulating a dataset. In a data matrix each case has a row to itself and each variable has a column.

A **vector** is a column or row of information from a data matrix. A vector is sometimes also referred to as a **field** of information.

A **cell** is a single unit of information in a data matrix.

In the example in Table 3.7 the dataset is arranged in a matrix. It is this form of arrangement and display of data that lies behind the formation of a spreadsheet using a computer. The cases (in this example, data relating to individual trade areas and countries) can be read across the rows and the variables (imports, exports and re-exports) are shown in successive columns.

In Table 3.8 information about the social nature of four pubs is given in a matrix.

It is worth noting that **matrix notation** is sometimes used as a short-hand way of indicating the information to be found in a particular cell of a specific data matrix. For

Table 3.7 UK imports from, and exports and re-exports to, various regional groups and countries, 1965 (current prices in £millions)

Region (case)	Variable		
	Imports	Exports	Re-exports
EFTA ^a	666.9	555.0	18.2
EEC ^b	994.7	904.8	75.4
Eastern Europe	222.4	125.2	3.7
Southern Europe and West Africa	202.9	195.0	4.4
Turkey and the Middle East	325.4	243.1	4.6
Rest of Africa	604.7	550.8	7.4
Asia	456.5	473.3	9.8
West Indies	98.7	97.1	1.8
Central and South America	289.7	157.7	3.0
Russia	118.8	46.9	0.5
Germany	265.4	255.0	30.4
Netherlands	270.8	193.1	9.9
Belgium	121.8	169.2	4.6
France	190.5	177.2	16.2
Republic of Ireland	170.4	175.8	9.9
India	128.3	114.1	2.3
Australia	219.5	281.4	3.0
New Zealand	208.2	125.0	1.1
United States	671.4	493.7	21.0
Canada	458.2	200.6	7.4
Argentina	71.5	26.8	0.7
South Africa	199.8	263.5	3.9

^a European Free Trade Association: Austria, Denmark (including Greenland), Liechtenstein, Norway, Portugal, Sweden and Switzerland.

^b European Economic Community: Belgium, France, West Germany, Italy, Luxembourg, Netherlands.

Source: B. R. Mitchell and H. G. Jones, *Second Abstract of British Historical Statistics* (London 1971), pp. 136–40.

History by Numbers

Table 3.8 The ratio of non-caps to caps and of seats to spittoons in the best room of four pubs in Bolton, 24–28 January, unknown year in the 1930s

Pub	Seats per spittoon	Non-caps to caps	Condition of spittoons
A	8	1.5	Dry
B	5	0.5	Wet
C	4	0.6	Wet
D	7	1.0	Slightly wet

Source: Mass Observation, *The Pub and the People: A Worktown Study* (London 1987 edn), p. 205.

matrix notation each matrix is given a letter of the alphabet, each row and column a number. In the example of Table 3.7 which comprises Data Matrix A, A1,2 denotes the first row, second column, that is, exports to the EFTA. A4,3 denotes the fourth row, third column, that is, re-exports to Southern Europe and West Africa. In Table 3.8 (Matrix B), B3,3 denotes the condition of spittoons in Pub C ('wet').

The historian beginning a project in quantitative analysis must decide which unit he or she will treat as cases and which as variables relative to those cases that comprise the study. This decision will be based upon the nature of the enquiry or analysis. In the case of Matrix A the focus of research is to be UK trade with specific trading partners. As a result, the countries and trading blocks are made the cases and exports, imports and re-exports are the variables. Had the focus been re-exports, different re-exported goods may have become the cases with country and trade block destinations forming the variables. Had the focus of enquiry been traded commodities, different sorts of goods may have been selected as the cases with relevant export and import quantities and/or values as the variables. Once a decision has been made on the basis of research priorities, the information can then be arranged in a way suitable for clarity of display and/or further statistical analysis.

The presentation of tables and figures

It is important to note at this point some important but simple rules about the professional presentation and layout of *all* tables (and figures).

1. All tables and figures must have a *title* that briefly describes their content, indicates the time period covered and the units of measurement used.
2. All tables and figures must have a *footnote* (or other formal reference) usually starting with the word 'Source:' giving details of the precise source or sources of the information which they contain. This is very important because it is necessary to give readers an opportunity to check figures or to seek further information about the derivation of figures from the sources quoted. Without the opportunity for readers to do this the statistics themselves and any analysis or

conclusions based upon them will always be open to question. Without detailed source notes, people will rightly think that the initial figures could have come from anywhere or have been manufactured and no-one will be interested in the analysis based upon them.

3. Column and row headings should be brief but self-explanatory with units of measurement clearly shown.
4. Vectors of data which are to be compared should be close together and derived statistics such as percentages and averages should be next to the figures to which they relate, either in brackets within the same cell or in an adjacent vector.

Table 3.9 arising from Hilary Doda's work on English sumptuary laws has all of these qualities.¹ Here the figures are derived from two different secondary sources that cover different social groups. Burkholder has covered a spectrum of society whilst the Hayward figures are for gentry only. They are thus kept separate in the table to avoid distorting the evidence. In this case the social class groups in the left are ordinal data because they are arranged in a hierarchy from lower status to higher status.

Table 3.9 Potential violations of sumptuary law, 1327–1553

	1327–1487 (Burkholder)	Clothing Bequests	Violations	1485–1553 (Hayward)	Clothing Bequests	Violations
Labourers	2	0	0			
Artisans	160	80	0–7			
Merchants	178	77	0–7			
Professionals	48	28	0–5			
Gentry	42	22	*	413	146	3
Nobles	23	12	*			

Note: * = unspecified.

Source: H. Doda (2014) “Saide monstrous hose”: compliance, transgression and English sumptuary law to 1533’, *Textile History*, 45 (2), 181.

Initial questions about the data

As soon as a historian or social scientist is faced with quantitative data it is a healthy sign if there are immediately many questions that she wishes to ask. How were the data collected and what errors might they contain? Is the same sort/quality of evidence available for each case (hence readily comparable) or is the evidence patchy? For what purpose was the data originally selected and recorded, and how does that purpose line up with our interests as historians? If, for example, Table 3.7 was to be used as a source for investigating the economic arguments for joining or opposing membership of the EEC in the late 1960s, the historian may well feel frustrated. The overlaps (between

History by Numbers

trading blocks and countries), as well as omissions and inconsistencies of the data collected from British official trade records, would be problematic. We might also wish to question whether a matrix for a single year would be representative. Was 1965 unusual in any way or was the UK consistently heavily tied to trade with the Commonwealth, EFTA, North America and South Africa before joining the EEC? Figures for other years would certainly be needed if the focus of research hinged upon the EEC entry debates. It would be necessary to get a sense of any growth or change in the direction of trade over time and a breakdown of trade by type of goods would be vital in assessing whether any losses which might occur from EEC entry could be substituted easily by trade within the EEC.

Similarly, we know that the material in Table 3.8 was gathered by volunteers organized by upper middle-class academics involved in the Mass Observation group and anxious to investigate the culture and leisure habits of the industrial working classes.² Should we be sceptical about their findings or the basis and validity of their research? Why were spittoons such a source of interest and were these academics correct in thinking that the level of spitting might tell us something about the 'social class' nature of the pub or the levels of chest congestion experienced by their customers? As historians, rather than as statisticians we should always be questioning the origins and reliability of the data for our purposes and asking what other sorts of data we would ideally need for our research. These questions are the essential prerequisite of quantitative analysis.

Apart from the data matrix, there are several other commonly used tools of descriptive statistics. Each shares with the matrix the attempt to impose order on information. We must bear in mind that the process of ordering and classification always results in the loss of some of the richness and detail of the original source. In addition decisions often need to be made about the rounding of measures, the selection of categories, and so on that can sometimes unwittingly or wittingly bias the results or introduce partiality or bias in the way that the results are displayed. These problems should always be appreciated and, where necessary, discussed as part of the research. The pros and cons of choosing particular categories, rather than others, and the likely loss or distortion arising from summary figures should be outlined either in the text or in a footnote.

With this in mind the rest of this chapter concentrates upon the benefit of the careful choice of descriptive statistics that will assist with immediate visual appreciation of the data, in summarizing large datasets and in providing the basis for further statistical analysis.

Grouping data in a frequency distribution

A **distribution** is a range of values for any one variable (for example, the range of values in a column vector). A **frequency** is the number of times any one value of the variable occurs. A **frequency distribution** is a tabulation that shows the frequency with which a particular variable occurs.

There are 3 types of frequency distribution:

- Simple;
- Grouped;
- Cumulative.

Each can be expressed in original units of measurement or in percentages. The choice of which frequency distribution to use depends upon which aspect of the data needs to be highlighted or examined.

The best way to illustrate the formation and use of frequency distributions is to use some further examples. Consider the data contained in Table 3.10 which is an extract from the return of convicts confined in Portland Prison in 1849.

Table 3.10 Return of convicts confined in Portland Prison, 1849

Name	Age (years)	Offence	Place of committal	Sentence (years)
James Hackett	21	Felony	Salford	7
John Taylor	20	Stealing a file and monies	Leicester	7
John Brown	20	Larceny (PC)	CC Court	7
James Barker	47	Stealing fowls, two indictments	Exeter	14
William Johnson	25	Setting fire to sacks of straw	Stafford	20
James Sweeney	58	Uttering counterfeit coin (PC)	Caernarvon	15
George Williams	21	Burglary (PC)	CC Court	10
Francis Best	35	Housebreaking, larceny	Worcester	15
John Henry	36	Uttering forged notes	Glasgow	20
Thomas Hartshorn	33	Robbery with violence	Liverpool	15
Samuel Laughton	22	Burglary, stealing silver spoons, etc.	Nottingham	14
Thomas Robinson	23	Burglary and theft, two indictments	Maidstone	14
Martin Stone	22	Horse stealing	Dorchester	15
Richard Ashford	58	Stealing 3lbs of pork	Exeter	10
John Dobson	28	Stealing money from the person (PC)	Stafford	14
Samuel Diggle	36	Burglary	Liverpool	15
George Goult	22	Robbery (PC)	Chelmsford	12
Robert Holder	23	Stealing from a dwelling £15 and pair of pistols	Portsmouth	15
Richard Jones	36	Warehouse breaking, and stealing malt and hops	Reading	15
Hugh King alias Cameron	36	Theft by housebreaking	Glasgow	14
Austin Montroe	34	Larceny in a dwelling to the value £5 (PC)	CC Court	15

Note: PC = previous conviction; CC Court = Central Criminal Court, London.

Source: Based on an extract from the Home Office, 8/102, The National Archives.

History by Numbers

We can quickly and clearly demonstrate features of the data such as sentence lengths, ages of prisoners, and geographical origins of prisoners by drawing up some frequency distributions.

In the simplest form we can add up the number of cases experiencing discrete sentence lengths. Discrete in statistical parlance means separate, discontinuous and referring to distinct objects, as in Table 3.11.

It is often clearer to group the data into particular bands as in Table 3.12. With just this short extract from the Portland Prison data this represents no major advantage over Table 3.11 because there are only six discrete categories of sentence. However, if the data were more extensive, with many different sentence lengths, the grouped distribution would represent a major advantage in terms of clarity in conveying the character of a lot of information very clearly and simply.

Table 3.11 Simple frequency distribution of sentence lengths (unrelated to type of crime) of prisoners in Portland Prison, 1849

<i>Sentence length (years)</i>	<i>Number of prisoners</i>
7	3
10	2
12	1
14	5
15	8
20	2
Total	21

Source: see Table 3.10.

Table 3.12 Grouped frequency distribution of sentence lengths (unrelated to type of crime) of prisoners in Portland Prison, 1849

<i>Sentence length (years)</i>	<i>Number of prisoners</i>
0–6	0
7–13	6
14–20	15
21–27	0
Total	21

Source: see Table 3.10.

In Table 3.13 information concerning prisoners' origins is given in a grouped frequency distribution. If left in the original discrete categories the frequency distribution would convey little more than Table 3.10. In grouped form Table 3.13 provides immediate clarity in conveying the distinctive regional pattern of prisoners' origins as indicated by place of committal. Table 3.13 also includes a percentage column. Percentages are often more useful than the original units in enabling one to see, at a glance, the shape of the

Table 3.13 Grouped frequency distribution of prisoners' origins as indicated by place of committal

UK area	<i>Prisoners</i>	
	number	per cent
London	3	14
Midlands	3	14
East	1	5
South	7	33
North	7	33
Total	21	

Source: see Table 3.10.

distribution. Percentages are also invaluable if we wish to compare the distribution of two or more datasets of different sizes (for example of different groups of Portland prisoners at different dates) or of different data (such as male and female prisoners).

Of course the regional pattern of prisoners' origins may not accurately be estimated from the places of committal not least because Central Criminal Court cases are likely to have come from further afield. In addition, of course, no account can be taken of the degree to which crime and committal took place away from a prisoner's normal place of residence nor can we allow for possibly very high geographical mobility of criminals. These are the sorts of points that it is important to add in order to qualify results or to draw attention to the weaknesses of any conclusions that you might draw.

Sometimes a **cumulative frequency distribution** is the most useful choice as in Tables 3.14 and 3.15 because it conveys an easier appreciation of the character of the data. Prisoners' ages are probably better expressed in this way than in a simple grouped distribution. Again percentages are especially useful when comparing the distribution of two datasets of differing size and composition. You will note that in the percentage frequency tables the percentages have been rounded to one decimal place. It is good practice to round to the nearest whole number or to one decimal place in most circumstances. Any greater precision would result in the sort of spurious accuracy that can get quantitative historical work a bad name.

Table 3.16 shows various frequency distributions for hearth tax payers in two Yorkshire townships in the 1660s. The hearth tax returns give the names of heads of households with the numbers of hearths on which they paid tax. Percentage frequency distributions enable comparisons to be made between the two townships.

Similarly, the land tax payers of the West Yorkshire township of Sowerby listed in Table 3.17 can be regrouped to provide a clearer indication of the distribution of landholdings than it is possible to see from the raw data itself: Table 3.18. The choice of class intervals here is of course dependent upon the researcher and in turn this should be determined by the level of detail required to address the research questions posed.

History by Numbers

Table 3.14 Cumulative grouped frequency distribution of prisoners' ages, Portland Prison, 1849

Age (years)	Number of prisoners
<20	0
<25	9
<30	11
<35	13
<40	18
<45	18
<50	19
<55	19
<60	21

Note: the sign ' $<$ ' means 'less than'.

Source: see Table 3.10.

Table 3.15 Percentage cumulative grouped frequency distribution of prisoners' ages, Portland Prison, 1849

Age (years)	Prisoners	
	number	per cent
≥50	2	9.5
≥40	3	14.3
≥30	10	47.6
≥20	21	100

Note: the sign ' \geq ' means 'greater than or equal to'.

Source: see Table 3.10.

The nominal data in the occupation field could also be grouped as the basis for asking questions about a possible relationship between occupation and value of landholding. (We use the occupational data later in a pie chart, Figure 3.13.)

Both the hearth tax and land tax returns have been the subject of extensive debate as to their accuracy and potential uses for the historian. The 1662 hearth tax return for Yorkshire included exempt households and is therefore more useful than other returns which did not. The accuracy and completeness with respect to names and numbers of householders is debatable but the hearth tax is a key source, especially for estimating wealth distribution and for demographic history. Many attempts have been made to estimate population totals from hearth tax figures by using multipliers to represent the average household size.³ The land tax returns hold similar pitfalls for the historian. Their accuracy appears to have varied greatly from one region to another and acreages are rarely given which is frustrating for historians wishing to use the returns to investigate the distribution of landholdings rather than levels of taxation. Because of differing land

Table 3.16 Analysis of hearth tax returns in the Yorkshire townships of Sowerby (including Soyland) and Calverley, 1664

Number of hearths	Household heads	
	number	per cent
<i>Sowerby</i>		
Exempt	140	30
0	3	1
1	185	39
2	72	15
3	29	6
4	27	6
5	10	2
6	2	<1
7	1	<1
8	1	<1
9	1	<1
Total	471	100
<i>Calverley</i>		
Exempt	43	34
0	0	0
1	58	46
2	18	14
3	6	5
4	1	1
5	:	:
14	1	1
Total	127	100

Source: Pat Hudson, 'Landholding and the organisation of textile manufacture in Yorkshire rural townships c. 1660–1810', in M. Berg (ed.), *Markets and Manufacture in Early Industrial Europe* (London 1991), p. 272.

Original source: Hearth tax returns, E179/210/393, 16 Charles II, Lady Day 1664, The National Archives.

values across the country (upon which taxation was generally based) it is very difficult to calculate acreage equivalents from the tax assessments. Furthermore, it is difficult to use the returns to look at wealth structures in terms of landownership because many people owned and rented land across several different land tax assessment boundaries. One must also be wary about studying change over time in landownership or occupancy from the returns because the land tax often went unrevised from year to year.⁴ The land tax, as a source for historians, is fraught with many difficulties and pitfalls but it is not unusual in this respect. Quantitative (as well as qualitative) evidence from all sources must be approached with caution before any attempt to analyse them is made.

Table 3.17 Land tax payers, Sowerby, West Yorkshire, 1782

Name	Land Tax paid			
	Occupation	£	s	d
John Batty	Clothier	1	3	6
Benjamin Bramley	Fuller	2	0	6
John Butterworth	Clothier		19	6
Abraham Clegg	Weaver		3	6
Thomas Cockcroft	Weaver		5	6
William Crossley	Clothier	1	8	0
John Derden	Merchant	5	0	0
Henry Dyson (3)	Yeoman	9	10	0
William Ellis	Yeoman	12	6	0
Eli Fielding	Weaver		2	6
Abe Gibson	Yeoman	5	0	0
John Gledhill	Clothier		15	6
David Greenwood	Clothier	1	13	6
John Greenwood	Miller	1	0	6
John Greenwood (3)	Merchant	12	0	0
Thomas Greenwood	Inn keeper		7	6
Cornelius Haigh	Fuller	2	0	6
John Hanson	Weaver		5	6
John Howarth	Clothier	1	10	0
Richard Hinscliffe	Clothier		17	6
Joshua Horton	Victualler		10	6
Wats Horton (24)	Gentleman	46	17	6
John Hoyle	Clothier		19	6
Richard Ingham	Clothier	1	4	0
John Irving	Woodcutter		2	6
John Lea (2)	Merchant	4	10	0
William Moore	Yeoman	2	5	0
Grace Ogden	Widow		2	6
Robert Parker (6)	Attorney	17	8	0
Danile Phillips	Clothier		5	0
Elizabeth Pimley	Widow		17	6
James Riley	Butcher	2	10	6
Joshua Riley	Clothier		7	6
Matthew Scott	Weaver		3	6
George Stansfield (15)	Merchant	30	2	6
Will Walker	Merchant	12	0	0
John Swain	Clothier		12	0
Ann Swain	Clothier	1	0	0
Matthew Tillotson	Weaver		10	0
Richard Thomas	Clerk	2	0	0
Sir John Deardon	Gentleman	7	10	0
John Walker	Weaver		3	6
John Walker	Clothier		15	0
Samuel Waterhouse	Merchant	8	0	0
John Whitaker	Clothier		19	6
Mary Whitworth	Weaver		4	6
Samuel Wood	Apothecary	2	0	0

Note: where a proprietor was liable to tax on more than one property the number of properties is given in brackets after the name.

Source: Land Tax returns, 1782, West Yorkshire Archive Service, Halifax, with hypothetical occupational data.

Table 3.18 Grouped and percentage grouped frequency distribution of land tax payers in Sowerby, West Yorkshire, 1782

<i>Tax paid</i>	<i>Number</i>	<i>Per cent</i>
< 5 s	7	15
5 s-< £1	15	32
£1-< £5	14	30
£5-< £10	5	11
£10-< £20	4	8
≥ £20	2	4
Total	47	100

Source: see Table 3.17.

A further example of a grouped frequency distribution is provided in Table 3.19 drawn from data in the very useful electronic source: Historical Directories of England and Wales (Special Collections Online, University of Leicester at: <http://specialcollections.le.ac.uk/cdm/landingpage/collection/p16445coll4>). A percentage column has also been included which gives a clearer immediate perception of the distribution. Nottingham was at this time concentrating upon the manufacture and sale of stockings with a clear separation of function visible between manufacturing and merchanting.

Table 3.5 was another example of percentage frequency data. In this case for two different years relating to the distribution of different fabrics in the wardrobes of Parisians of different social and occupational groupings. Taken out of the context of research by W. H. Sewell on which they are based, the figures raise many questions particularly about the apparent substitution of cotton for linen between the two dates across all social groups and the apparent substitution of cotton for woollen clothing amongst wage earners. The figures prompt questions about whether the

Table 3.19 Manufacturers and tradesmen in Nottingham, 1783

<i>Occupation</i>	<i>Number</i>	<i>Per cent</i>
Manufacturers of hose	65	43
Manufacturers of other textiles	16	11
Manufacturers of other goods	7	5
Textile merchants	21	14
Sellers of other goods	36	24
Merchant-manufacturers	5	3
Total	150	100

Source: Data from *Baileys' Western & Midland Directory the Borough of Nottingham* (Nottinghamshire 1783) at: <http://specialcollections.le.ac.uk/cdm/compoundobject/collection/p16445coll4/id/112445/rec/26> (accessed 20 September 2015).

History by Numbers

various changes represent improvements or deterioration in wealth or living standards, changes in taste or fashion or the availability of new goods. Information regarding relative prices of various fabrics and some analysis of their practical or fashionable appeal would be needed in order to take the analysis further. Above all however, and particularly because the information for each social class was likely derived from different sorts of primary sources (thence not directly comparable), the keen historian should be questioning the reliability of the information in relation to Sewell's argument or purpose.

There are four important things to remember about the formation and presentation of frequency distributions:

1. Each needs both a correct heading indicating its content and/or purpose and a footnote (or other form of reference) indicating the source of the information.
2. Columns must always have an appropriate heading.
3. Simple and grouped frequency distributions should always include a total figure (this is useful anyway as a check that all the cases have been included).
4. Most importantly, grouped distributions must be sure to have *no overlap* between the categories or class intervals.

Bar charts

Bar charts and histograms are both ways of displaying the sort of data collected in frequency distributions. Bar charts and histograms present the information in the form of a figure rather than a table. Their advantage is that they often give a clearer and more immediate visual representation of the data than that given in a frequency table.

Bar charts can be used for nominal, ordinal or interval data. In them the bars are normally separated from one another and the variables can appear in any order. The length (or height) of the bars is proportional to the observed frequencies. The width and area of the bars can vary and is not important. Examples of bar charts are given in Figures 3.1–3.5. It is common to display the frequencies of different variables side by side as in Figures 3.1 and 3.2. In the former, the arrangement of the data into a bar chart makes it clear that inheritance was partly determined by gender and that wives, if they did not inherit the whole estate as a widow, would be most likely to receive household goods and least likely to receive a business or investments. In Figure 3.2 the bar charts demonstrate not only that the number of Gladstone's parliamentary speeches varied greatly from one decade to the next but that domestic affairs dominated whilst Ireland was the focus of many speeches between 1860 and the mid-1890s.

In Figure 3.3 change over time is incorporated by showing a number of bar charts each with different variables side by side. This is a good example of the applicability of bar charts to the analysis of qualitative data and literary history. The bar charts

Arranging, Rearranging and Displaying Data

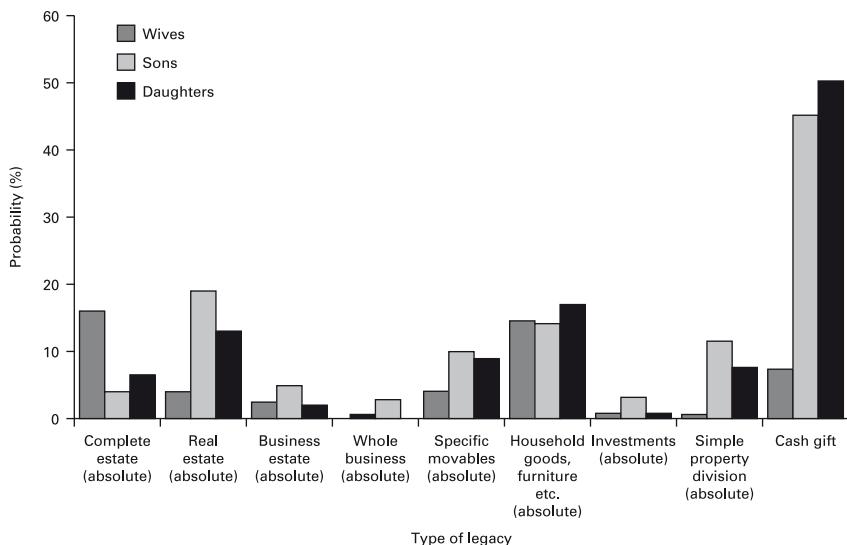


Figure 3.1 Probability of wives and children receiving different types of legacy as an absolute gift, Consistory Court of Chester, 1800–1857.

Source: Alastair Owens (2001) 'Property, gender and the life course: inheritance and family welfare provision in early nineteenth-century England', *Social History*, 26 (3), 299–317, p. 311.

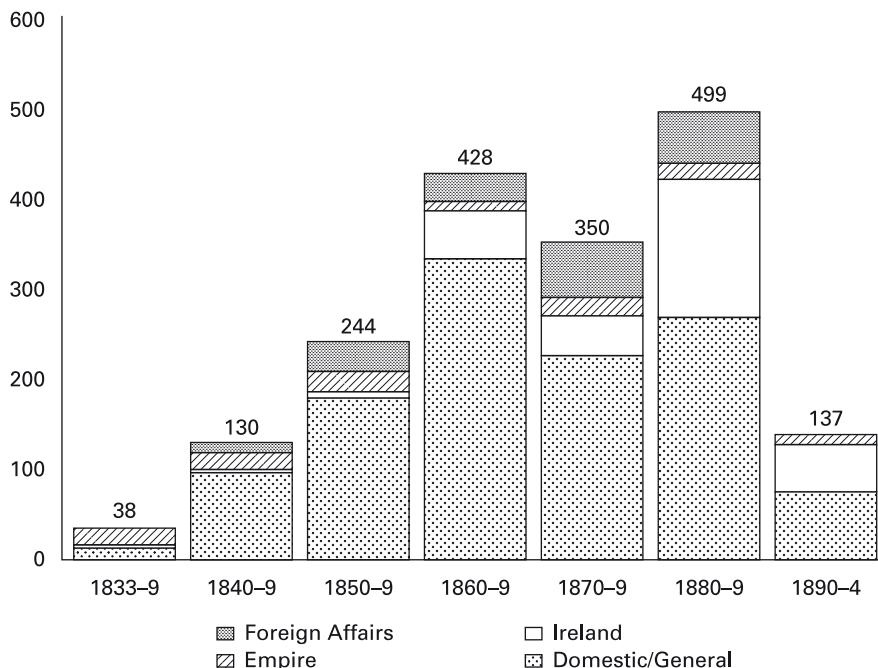


Figure 3.2 Subject matter of Gladstone's speeches in the House of Commons, by decade, 1833–1894.

Source: Joseph S. Meisel (2000) 'Words by the numbers: a quantitative analysis and comparison of the oratorical careers of William Ewart Gladstone and Winston Spencer Churchill', *Historical Research*, 73 (182), 262–295, p. 273.

History by Numbers

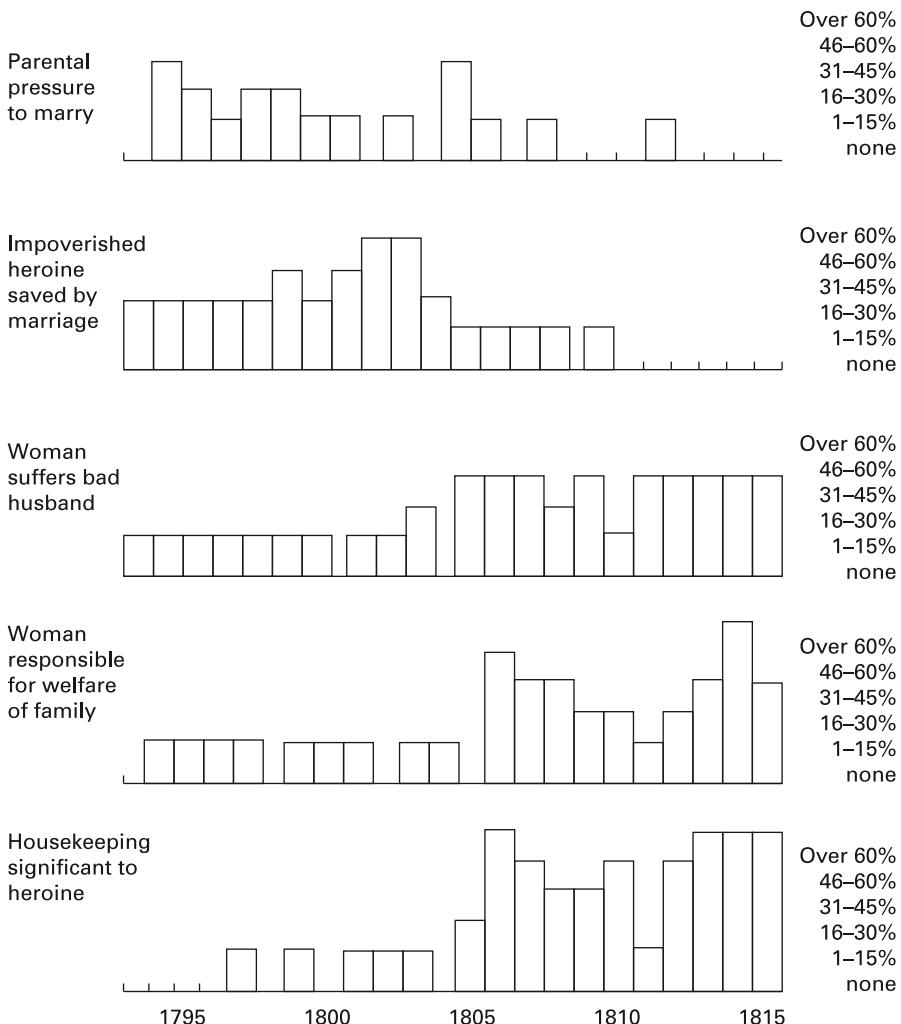


Figure 3.3 Plot analysis bar charts from *The Lady's Magazine*, 1793–1815.

Source: Edward Copeland (1995) *Women Writing About Money: Women's Fiction in England, 1790–1820*, Cambridge: Cambridge University Press, p. 63.

make it clear that, in plot lines of stories published in *The Lady's Magazine*, parental pressure to marry and impoverished women saved by marriage both declined as the Napoleonic Wars came to a close but that women increased their roles in the welfare of the family and in concern about housekeeping at the same time, and were more likely to suffer a bad husband. How much the historian reads in to these plot changes and sees them as a reflection of real concerns or conditions in society (and the social classes that might have been involved) is of course entirely a matter of historical judgement.

Each bar of a bar chart can also be divided to represent further features of the data as in Figure 3.4 (a component bar chart where the components are expressed as a proportion of the whole).

Figure 3.4 shows that the number of travelling salesmen working for the McVitie biscuit company expanded in the late 1920s and early 1930s and that most were better paid than in the past. The bars in a bar chart can be arranged vertically or horizontally (as in Figure 3.4). The latter is generally preferred when hierarchical data is being displayed as in Figure 3.5.

Figure 3.6a contains a mass of raw data that is difficult to interpret in that form. Place of birth data for Glasgow has been extracted from the table for the bar charts in Figure 3.6b. This is a good indication of the level of clarity and efficiency of communication that can be achieved with the aid of a computer graphics package, allowing the production of four bar charts in three dimensions side by side for comparative purposes, once the Glasgow data has been entered into an electronic database.

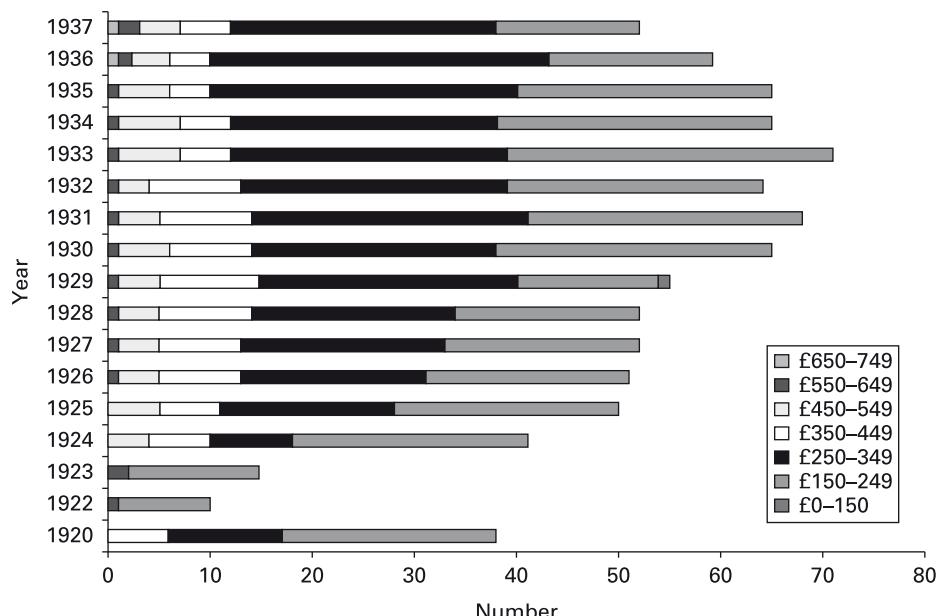


Figure 3.4 Distribution of McVitie travellers by income, 1920–1937.

Source: Michael French (2005) 'Commercials, careers, and culture: travelling salesmen in Britain, 1890s–1930s', *Economic History Review*, 58 (2), 352–377, p. 361.

History by Numbers

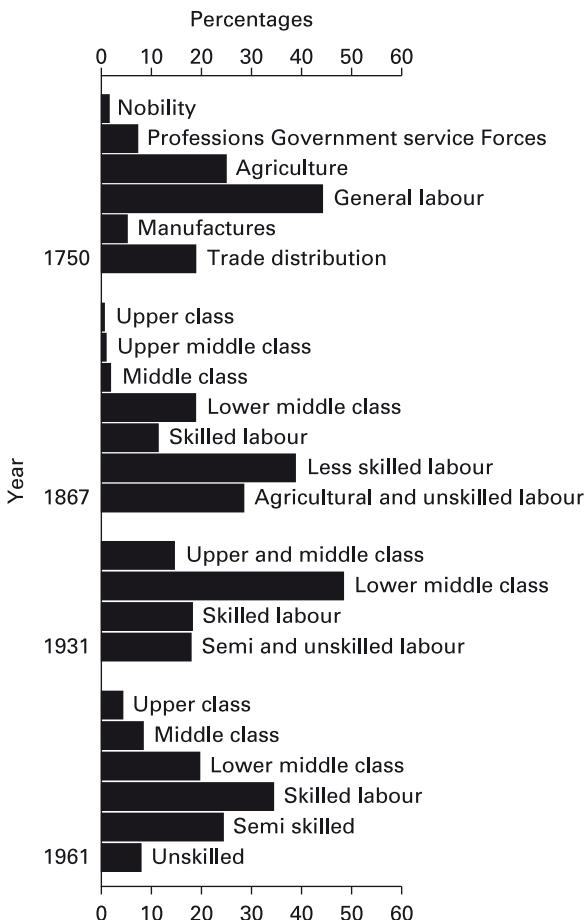


Figure 3.5 Bar charts of class structure (%), 1750–1961.

Source: Based upon E. J. Hobsbawm, *Industry and Empire* (London 1968), p. 303.

Figure 3.6a Facsimile of 1881 census, Glasgow Govan and Galashiels

Source: E. Mawdsley and T. Munck, *Computing for Historians* (Manchester 1993), p. 132.

History by Numbers

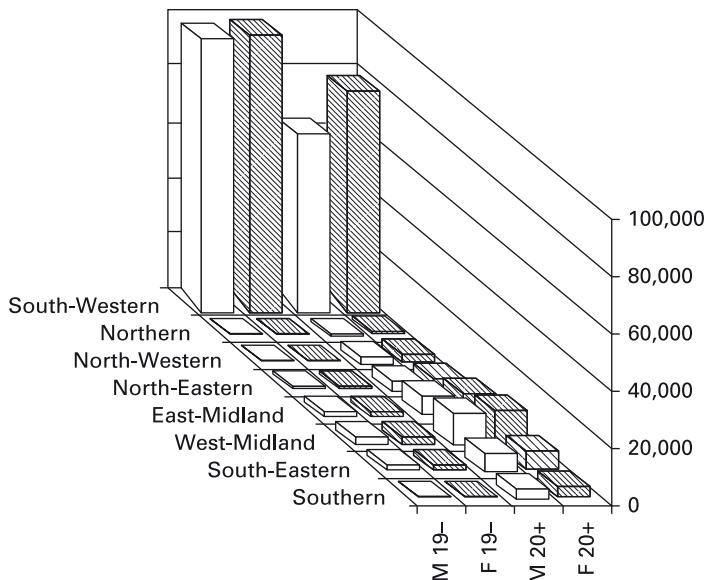


Figure 3.6b Bar chart showing Glasgow population (municipal burgh) by place of birth, 1881.
Source: E. Mawdsley and T. Munck, *Computing for Historians* (Manchester 1993), p. 134.

Histograms

Interval data are more usually represented by a **histogram**. A histogram is a diagrammatic representation of a frequency distribution consisting of a series of rectangles or bars with a width proportional to the class interval and an area proportional to the frequency.

In a histogram data values are continuous rather than discrete and appear next to one another, normally along the horizontal axis. The width of each bar of a histogram is the same (providing that the data values or range of values for each bar is the same) though this can be varied if class intervals are of unequal size. The area covered by the bars (*and* the height of the bars, if the class intervals are equal in size) is always proportional to the frequency being represented.

The information about McVitie traveller income data (Figure 3.4) could be reconfigured into a histogram to show the number in each income band in any one year. Figure 3.7 demonstrates this for the two years 1929 and 1936, side by side. The likely impact of the 1929 crash upon the buoyancy of incomes is obvious.

Further examples of a histogram are given in Figures 3.8 and 3.9. Figure 3.8 has the additional feature of being placed on its side with two histograms back to back allowing immediate comparison between male and female populations. Using the facilities provided by a computer software package, histograms, like bar charts, can be drawn directly from a data matrix or spreadsheet using the appropriate software commands. If drawn in three dimensions the volume as well as the height of bars is proportional to the frequencies represented which gives further visual clarity to the data.

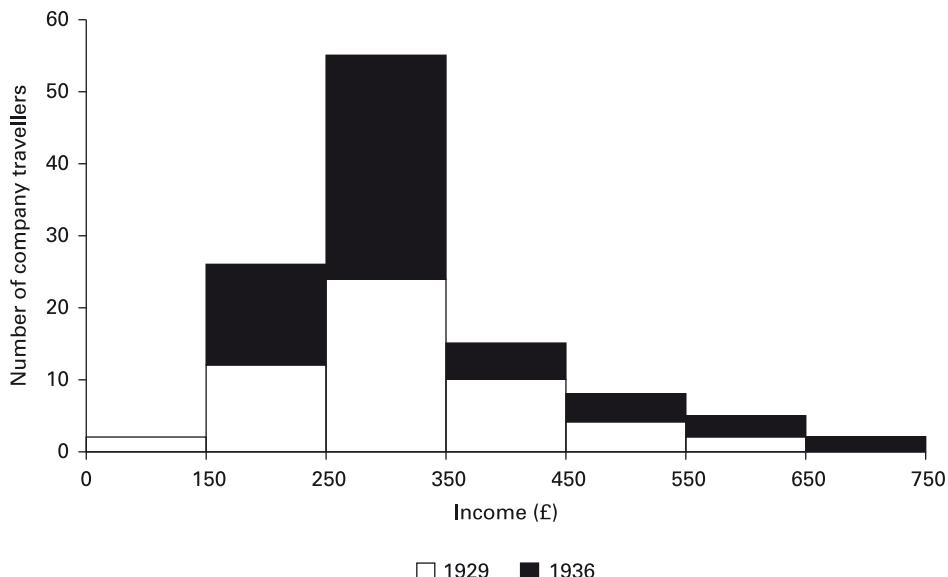


Figure 3.7 Income distribution of McVitie travellers, 1929 and 1936.

Source: Based on Michael French (2005) 'Commercials, careers, and culture: travelling salesmen in Britain, 1890s–1930s', *Economic History Review*, 58 (2), 352–377, p. 361.

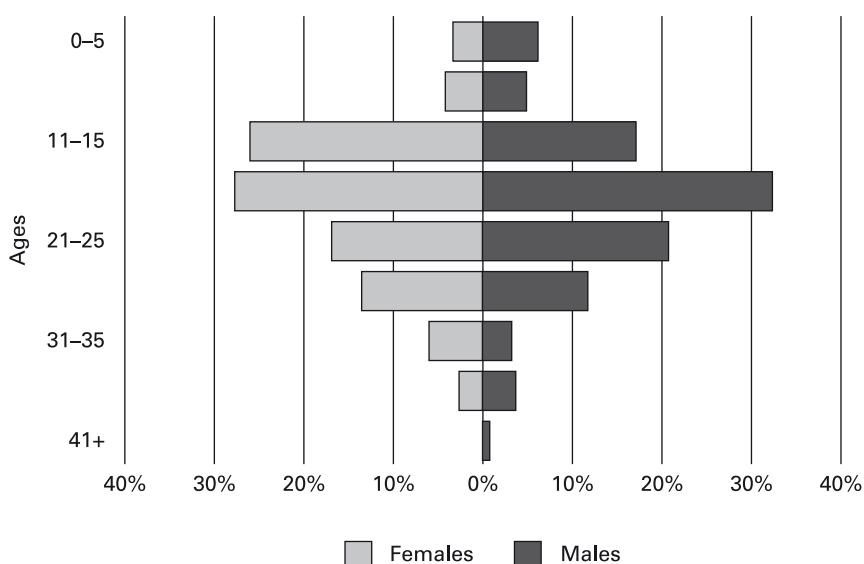


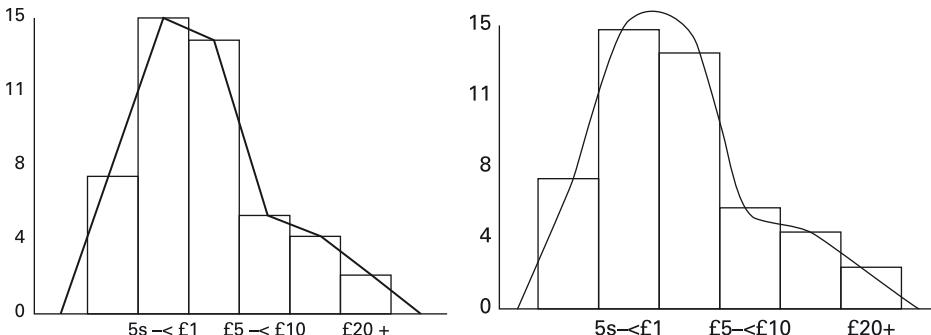
Figure 3.8 The age distribution of slaves advertised for sale, Boston, 1720–1781.

Source: Based on R. E. Desrochers Jr (2002) 'Slave-for-sale advertisements and slavery in Massachusetts, 1704–1781', *William and Mary Quarterly*, 3rd series, 59 (3), Table 4, p. 636.

History by Numbers

Sometimes the midpoints of the tops of histogram blocks are joined by straight lines to form a **frequency polygon** as in Figure 3.9a. Sometimes the histogram/polygon is smoothed into a **frequency curve** by using much smaller class intervals as in Figure 3.9b. Note that the area under the 'curve' remains the same as the area covered by the histogram.

The purpose of frequency polygons and frequency curves is to provide a clearer visual picture of the character of the frequency distribution. They also have the advantage over the histogram that several can be plotted on the same axis making their comparison much easier. In Figure 3.10 frequency polygons of mean annual mortality from tuberculosis, for different time periods, are displayed on the same axes for immediate visual comparison.



Figures 3.9a and 3.9b Frequency polygon of land tax payers in Sowerby, West Yorkshire, 1782.

Source: See Table 3.18.

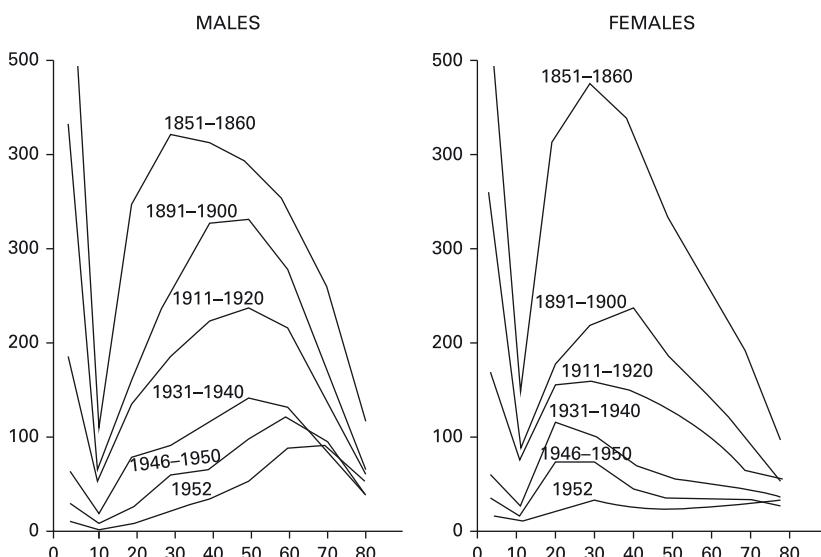


Figure 3.10 Frequency polygons of male and female mean annual mortality from tuberculosis at different ages, in selected periods.

Note: x-axis shows age (years) and y-axis shows mean annual mortality per 100,000.

Source: A. Hardy (2003) 'Reframing disease: changing perceptions of tuberculosis in England and Wales, 1938–1970', *Historical Research*, 76 (194), p. 550.

Important points:

1. In bar charts and histograms the variables or frequencies represented by the bars must be labelled clearly at the base of the bars themselves or by shading the bars and providing a key.
2. The vertical and horizontal axes must be clearly labelled.

Pie charts and pyramid charts

These provide a clear representation of the proportions of different categories or values found in a dataset. A **pie chart** is a circle or shallow cylinder divided into sectors to represent each item or variable. The sectors are exactly proportional to the distribution of the data. Each sector is normally shaded and labelled with a percentage whilst a key is also provided to identify the units or cases represented by the sectors. Figure 3.11 is a pie chart showing the proportions of UK re-exports by geographical or trading area in 1965. It is derived from the data in Table 3.7. The size of the sectors is calculated and measured by dividing the 360 degrees of a circle in the same proportions as that of the data. For example, the size of the sector for EFTA was calculated in the following way:

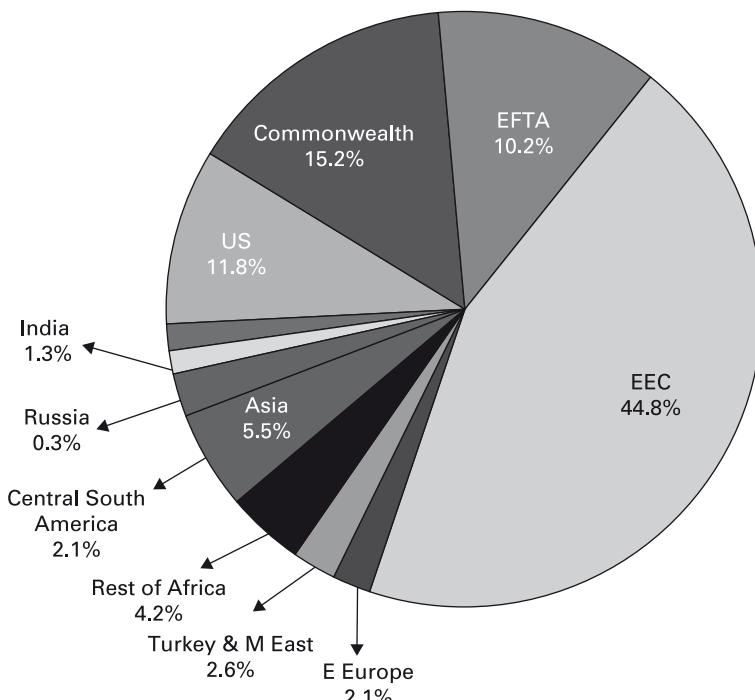


Figure 3.11 Pie chart showing proportions of UK re-exports by geographical or trading area, 1965.

Source: see Table 3.7.

History by Numbers

$$\begin{aligned}
 \text{angle } \alpha &= \frac{\text{total EFTA re-exports}}{\text{total re-exports, excluding double counting}} \times 360^\circ \\
 &= \frac{\text{£18.2 million}}{\text{£160 million}} \times 360^\circ \\
 &= 41^\circ \text{ (to nearest degree)}
 \end{aligned}$$

The sector for EFTA is therefore drawn so that the angle at the centre of the circle is 41 degrees. Do note that in this example care has been taken not to overlap or double count the re-export figures by including trade blocks in the pie chart alongside countries that are part of those trade blocks.

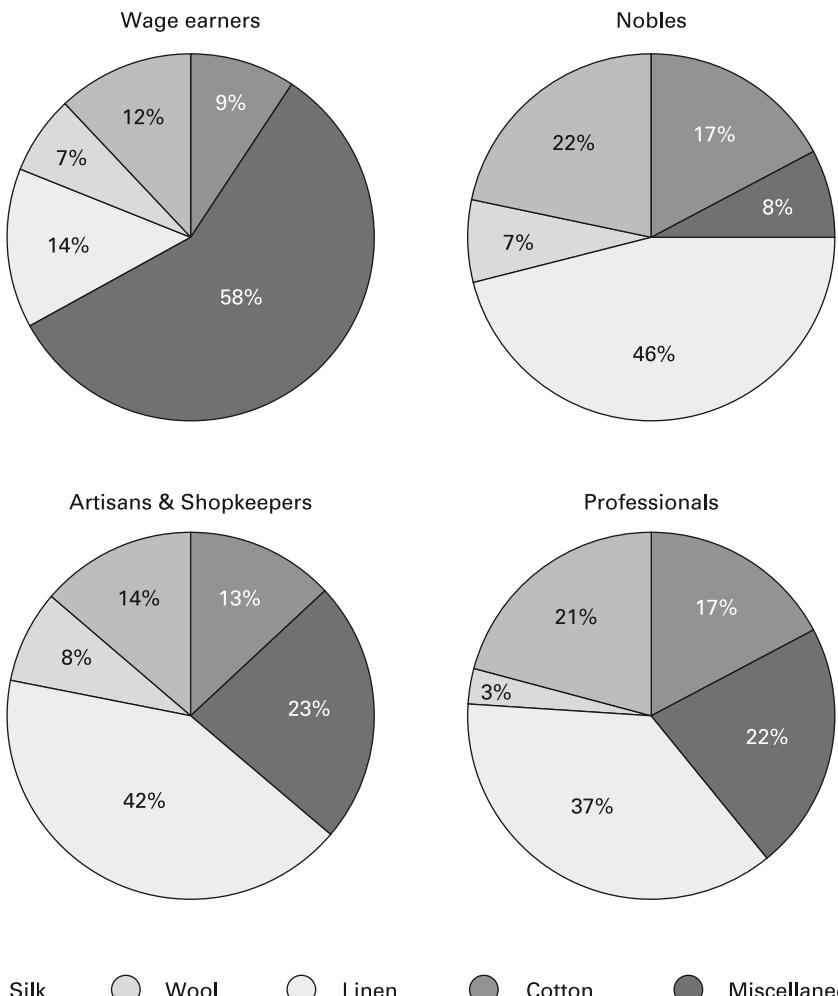


Figure 3.12 Pie charts showing the proportions of different types of fabric in the wardrobes of nobles, professionals, artisans and shopkeepers/wage earners in Paris in 1700.
Source: see Table 3.5.

In pre-computer times pie charts were drawn using a protractor or compass. Nowadays computer software can process the information from a spreadsheet or database and produce pie charts from a mass of data very easily. Often the software produces pie charts in three dimensions which can give further visual impact to the data display. Pie charts can be used side by side to give a rapid indication either of change over time or of comparative data as in Figure 3.11. From Figure 3.11 it is easy to see the advantage of comparative pie charts over the display of this data in a table. The pie charts in Figure 3.12 are based on the information in Table 3.5 for 1700. You may wish to test your skills by inputting the data for 1789 from the same table into a simple spreadsheet and drawing the comparative pie charts for that year also.

Figure 3.13 shows the occupational data from Table 3.17 in the form of a pie chart that gives a useful indication of land held by the main occupational groups in Sowerby.⁵

In the example in Figure 3.13 the main occupational groups are weavers, clothiers and merchants. Professionals include apothecary (1), attorney (1), and clerk (1). ‘Others’ include the butcher, innkeeper, miller, victualler, woodcutter, and widows (2). Weavers (8) are also included in this group because the amount of tax they paid in total amounted only to £2 1s. We know from other sources that many weavers and other textile workers in the township held no taxable land. Sowerby’s land ownership profile is typical of a textile putting out community where there was a fairly land-rich group of putting out merchants and clothier farmers who in turn employed wage dependent spinners and weavers. Spinners have a low profile in this sort of tax evidence (except where they are substantial widows) because so few women owned land in their own right and spinners were mostly drawn from the wives and daughters of farmers or other textile workers.

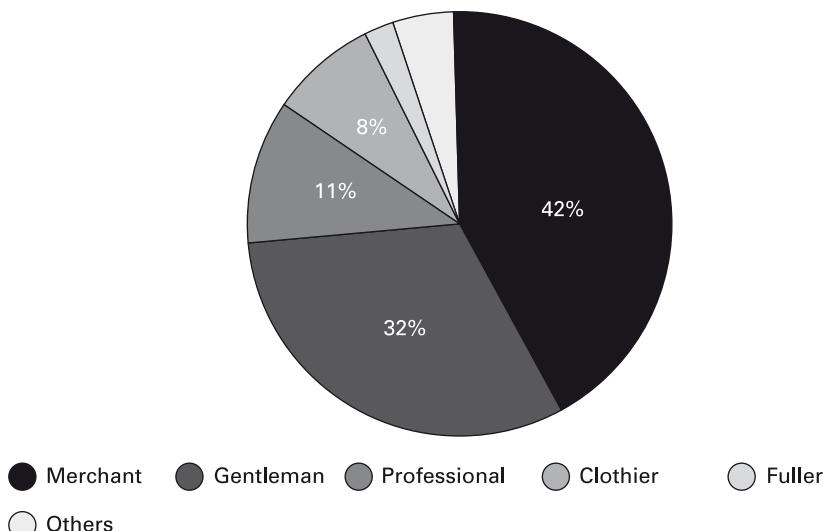


Figure 3.13 Landholdings in Sowerby by occupational group, 1782.

Source: see Table 3.17.

History by Numbers

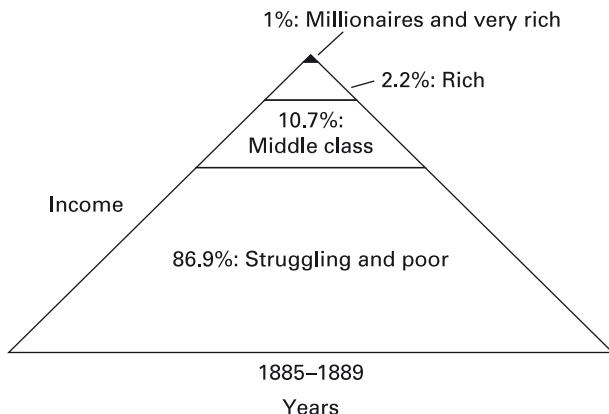


Figure 3.14 Pyramid chart of the Victorian rich and poor.

Source: Based upon E. J. Hobsbawm, *Industry and Empire* (London 1968), p. 308.

Pie charts are useful where there are relatively few variables which make up proportions of the whole, where it is more important to convey the proportions than the numerical values and where a strong visual emphasis is required.

Where data proportions reflect a marked hierarchy, a triangular chart is sometimes used with the same principle as the pie chart. This is called a **pyramid chart**. An example is given in Figure 3.14. Just as in a pie chart the area represented by the sections of the triangle are drawn so that they are proportional to the data distribution. In this case this is done manually using the formula for the area of a triangle (that is, half base-length multiplied by the perpendicular height). More easily nowadays it is done by feeding the data into a computer software package and giving the appropriate instructions.

Graphs: time series

Line graphs are especially useful for displaying **time series** data (that is, data that varies over time). They are therefore particularly common in historical use. They have a wide variety of other functions particularly where it is desired to represent the relationship between the movement of two or more variables.

In a time series graph, time is normally measured on the horizontal axis (usually referred to as the *x*-axis) which is marked out in months, years or another unit of time. The movement of one or more variables can then be drawn with respect to a scaled vertical axis (or *y*-axis). If more than one variable is depicted, a key or legend must be provided to indicate clearly which line represents which variable. Figure 3.15 shows the yearly profit rates for four firms in the worsted industry in the nineteenth century, which I researched in the early 1980s. Three firms ran at a loss in some years so I placed the horizontal axis accordingly. It does not have to be at the bottom of the graph. The vertical axis of a graph is also moveable if it is necessary to depict negative values. Displaying the

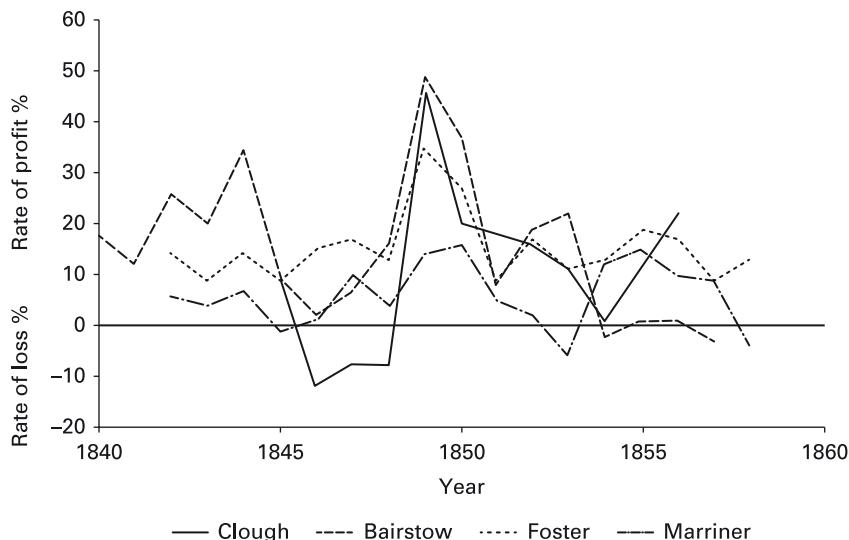


Figure 3.15 Profit rates in the worsted industry, 1840–1858.

Source: Pat Hudson, *The Genesis of Industrial Capital* (Cambridge 1986), p. 239.

experience of just a few firms in the industry at the time may of course be misleading. The available accounts may not be representative of the Yorkshire textile sector as a whole and indeed, it is often the case in business history that larger and certainly more successful firms are most likely to have generated records that have survived. One needs to keep a historian's keen eye in evaluating this data, as well as having an understanding of how to display the information.

Figure 3.16 shows the impact of smallpox on average height by age. The (horizontal) x-axis is placed in the centre (vertically) allowing negative and positive values on either side. Confidence intervals are calculated here to reflect the reliability of the data given the size of the sample (by comparing what was found with the distribution of heights occurring in a 'normal' population). Sampling and confidence testing are fully explained in Chapter 7. To decide whether or not the data represents wider experience of the impact of smallpox, we would need to learn about the representative nature of the convict data. We would also benefit from knowing what modern medical science can tell us about the impact of stunting diseases, experienced at various ages in childhood, upon adult height. The Oxley article is the subject of an exercise on p. 236.

Sometimes the different variables that we wish to depict on the same time series graph for comparative purposes have a different unit of measurement. In this case it is necessary to provide two or more separate scales on the same vertical axis or to have a second vertical axis with a separate scale or scales on the right hand side of the figure, as in Figure 3.17. Here it is possible to see the possible impact of fan introduction on productivity measured in three different ways, all on the same graph.

If different variables have the same unit of measure but one is very much smaller than the others, the smaller measure may be expressed in a multiple so that lines can be

History by Numbers

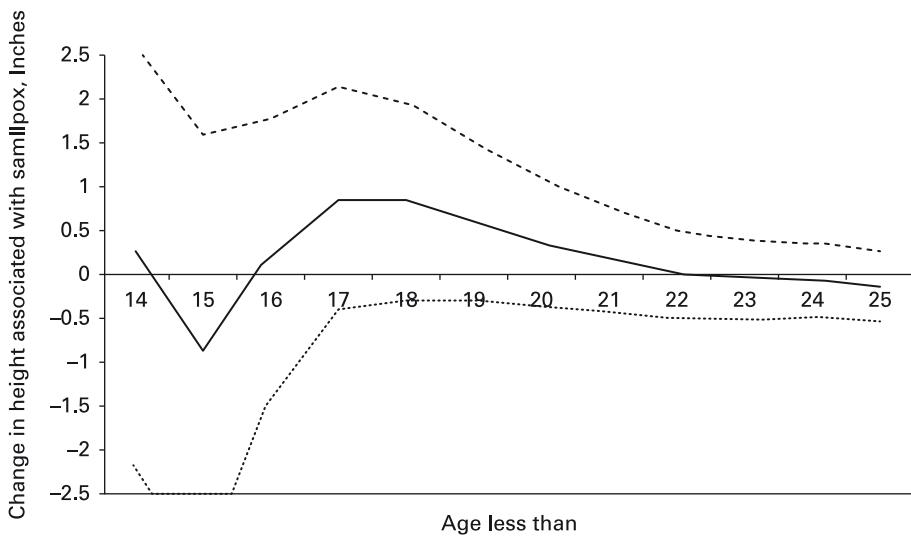


Figure 3.16 Impact of smallpox on average height, by age.

Source: D. Oxley (2006) “Pitted but not pitied” or, does smallpox make you small?, *Economic History Review*, 59 (3), 619.

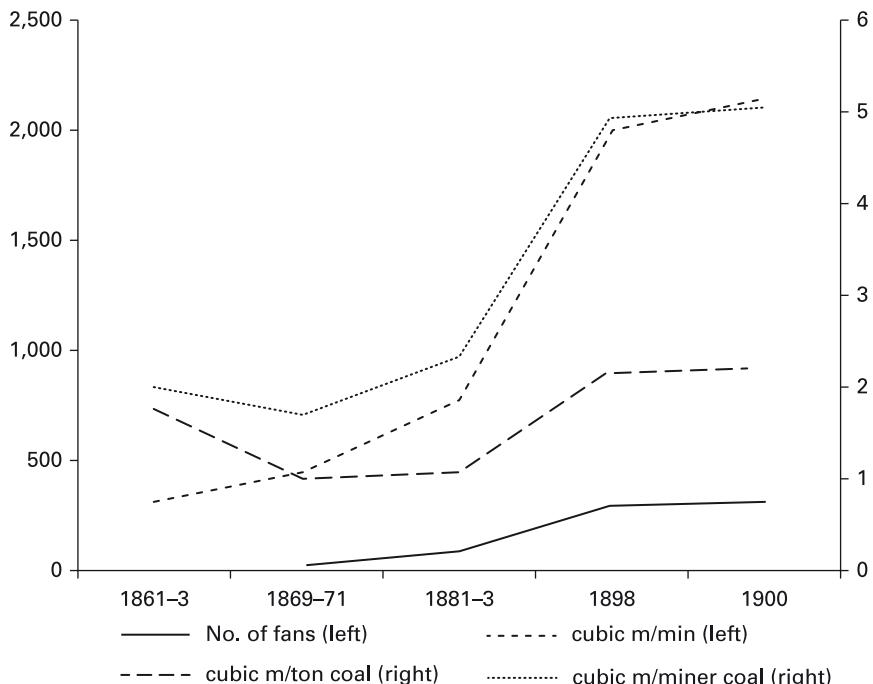


Figure 3.17 Fan introduction and diffusion in coal mines through the Ruhr district, 1861–1900.

Source: J. E. Murray and J. Silvestre (2015) ‘Small scale technologies and European coal mine safety, 1850–1900’, *Economic History Review*, 68 (3), 896, in turn based on U. Burghardt, *Die Mechanisierung des Ruhrbergbaus 1890–1930* (Munich, 1995).

compared on the same graph or, as in the case of Figures 3.17 and 3.18 (the latter drawn from sociopolitical history), a second scale can be provided on a second *y*-axis. The study portrayed in Figure 3.18 is the subject of an exercise at the end of Chapter 4.

If the variables to be depicted grow slowly at first but then accelerate rapidly a **log scale** may be used on the vertical axis to create a **semi-logarithmic graph**. A semi-logarithmic graph is one in which the measures on the vertical axis decrease successively at the upper end of the range of values because they are expressed in logarithms rather than the original units. The curve is thus made to fit on a page and be easily visible, as in Figure 3.19. The purpose of the semi-logarithmic graph is to show the rate of change of data rather than changes in the actual amounts: the slope of the curve indicates the rate of change because each point shows the percentage change from the last point.

Alternatively, if the variable exhibits exponential growth, the vertical scale can increase in units that are successively the square roots of the previous unit. This is termed an **exponential scale**. Both logarithmic and exponential scales have the same purpose in allowing the graphing of rapidly growing variables of different kinds.⁶

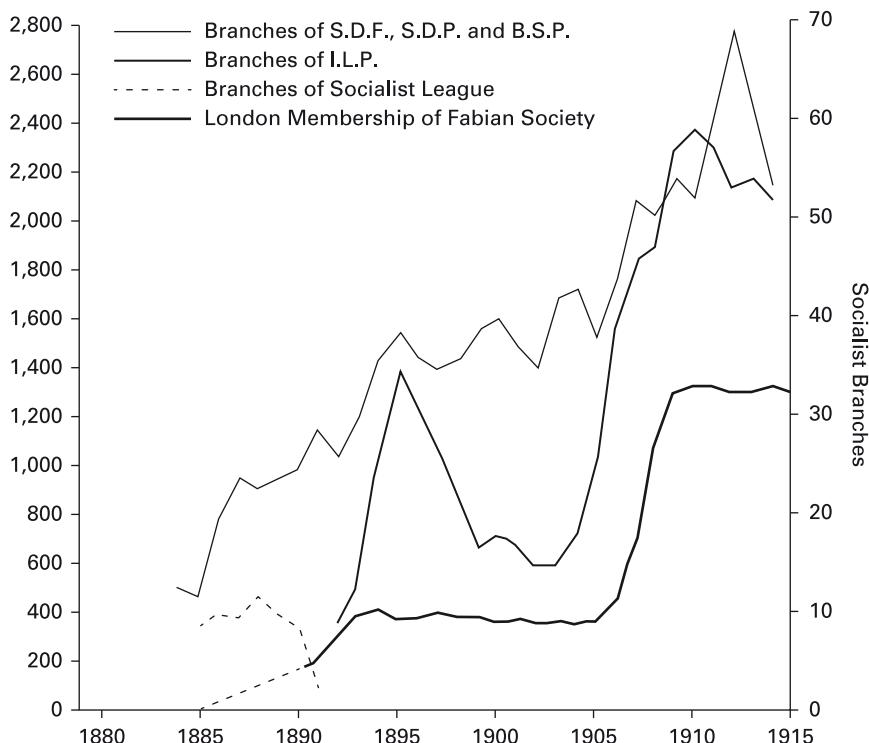


Figure 3.18 Strength of Social Democratic Federation, Fabian Society, Socialist League and Independent Labour Party in London, 1880s to 1915.

Source: D. M. Young (2005) 'Social democratic federation membership in London', *Historical Research*, 78 (201), 356.

History by Numbers

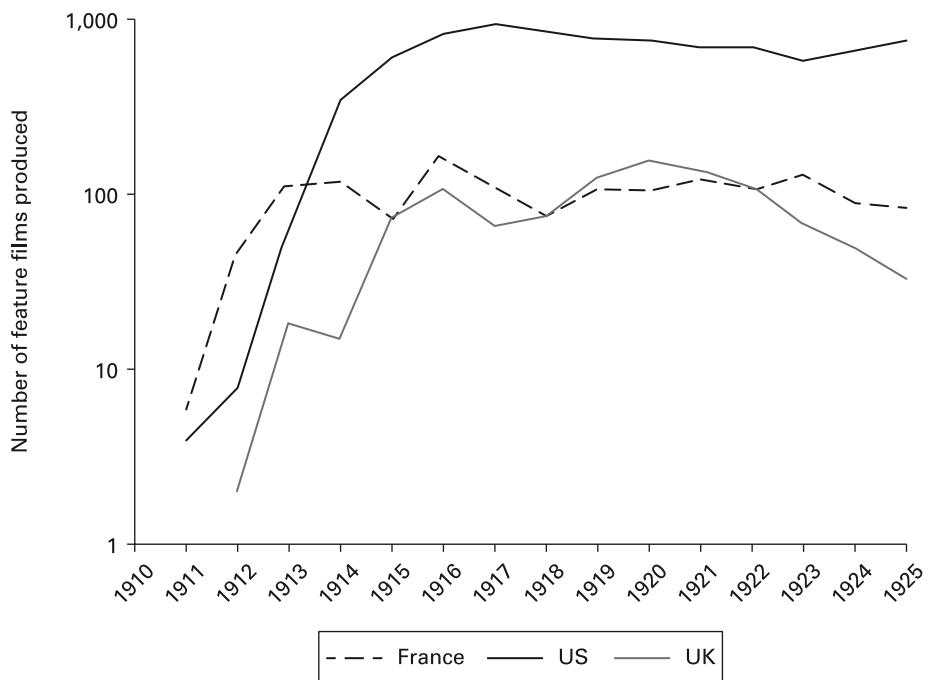


Figure 3.19 Number of feature films produced in UK, France and USA, 1911–1925.

Source: G. Bakker (2005) ‘The decline and fall of the European film industry; sunk costs, market sizes, and market structure, 1890–1927’, *Economic History Review*, 58 (2), 315.

One must, of course, be on the lookout for vertical scales of this kind because at first sight they may give a misleading impression of the rate of growth of the data. Figure 3.19 shows the growth of the early film industry in three countries. Growth in France appears to lag by about a year but the growth rate experience is similar in all three cases. As with most new technologies and cultural innovations, there is a common pattern of rapid initial take up followed by a levelling off of growth necessitating a logarithmic or semi-logarithmic scale to enable the data to be fitted on the figure and on the page.

Other graphs using independent and causal variables

Graphs have many other uses apart from time series. A common use is in graphing **dependent and independent variables**. In such a graph which is designed to indicate the relationship between the movement of two variables with respect to one another (for example height and weight of army recruits, exports and imports), it is not crucial which axis is used for which variable. However if one is analysing the movement of two variables whereby one is suspected as a prime cause of the movement of the other (for example, the price of corn and incidents of rural protest or the movement of real wages and meat

consumption) the causal variable or so-called **independent variable** is normally located on the horizontal axis with the **dependent variable** on the vertical axis. Thus in the first example, where high food prices may have been partly responsible for social unrest, corn prices would generally be placed on the horizontal axis and number of protest incidents on the vertical axis. Real wage movements (as a possible cause of variation in the consumption of meat) would be on the horizontal axis with meat consumption on the vertical scale. These so-called **scatter graphs** are more fully explored in Chapter 6 but Figure 3.20 gives one example. Matthew J. Hill here explores 'Love' in times of economic expansion and depression by considering the relationship between variation of Gross Domestic Product (GDP), that is, the output of the US economy (a proxy for economic circumstances and the putative causal variable) with the marriage rate per 1,000 people (the dependent variable). The latter is plotted against the former. Because it is possible to discern an upward (positive) slope in the plotted points, the suggestion is that the marriage rate is related in a positive way to economic conditions. Whether economic conditions (especially as measured somewhat indirectly by GDP per capita) is a determining factor in the variation of marriage rates and in determining marriage decisions is another question and one where further evidence and the historian's judgement will be needed.

Apart from questioning the appropriateness of studying love via the marriage rate, several things might be noted about this graph. First GDP is necessarily given a log scale

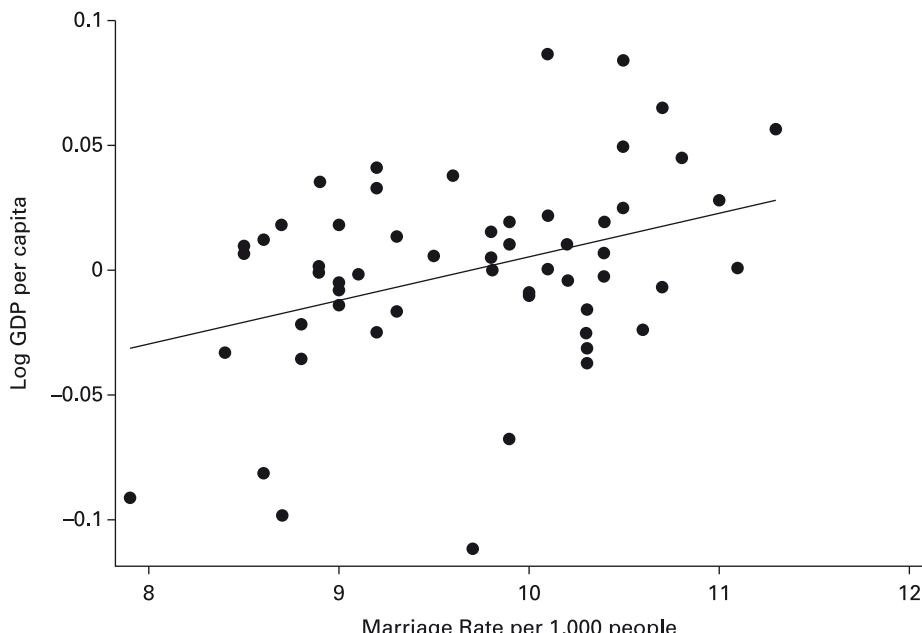


Figure 3.20 The relationship between marriage rates and detrended GDP per capita in the United States, 1887–1960 (First World War and Second World War excluded).

Source: M. Hill (2015) 'Love in the time of the depression: the effect of economic conditions on marriage in the Great Depression,' *Journal of Economic History*, 75 (1), 166.

History by Numbers

because the range of GDP per capita variation over the period would otherwise be too great to fit the page. Second, the GDP figures have been detrended to leave only the cyclical fluctuations. (For more on the pros and cons of doing this, and how it is done, see Chapter 5.) One might certainly question the wisdom of detrending the GDP series for the purposes of this research. Finally, and perhaps less controversially, the two war periods have been excluded because of their exceptional nature in influencing both GDP and the marriage rate and perhaps therefore disturbing the statistical relationship between the two variables that is visible over the long term.

Graphs can also be used to represent more novel or complex relationships. Figure 3.21 is one of many interesting graphs appearing in K. D. M. Snell's *Annals of the Labouring Poor* (1985). It shows an unusual use of graphical representation to demonstrate seasonal variation, in this case in female unemployment before and after enclosure. This excellent use of graphical representation was very effective in supporting Snell's analysis of the impact of enclosure upon women's work in the countryside although his assumption that the timing of Settlement Examinations and Certificates would so accurately reflect the timing of unemployment perhaps needs questioning.⁷

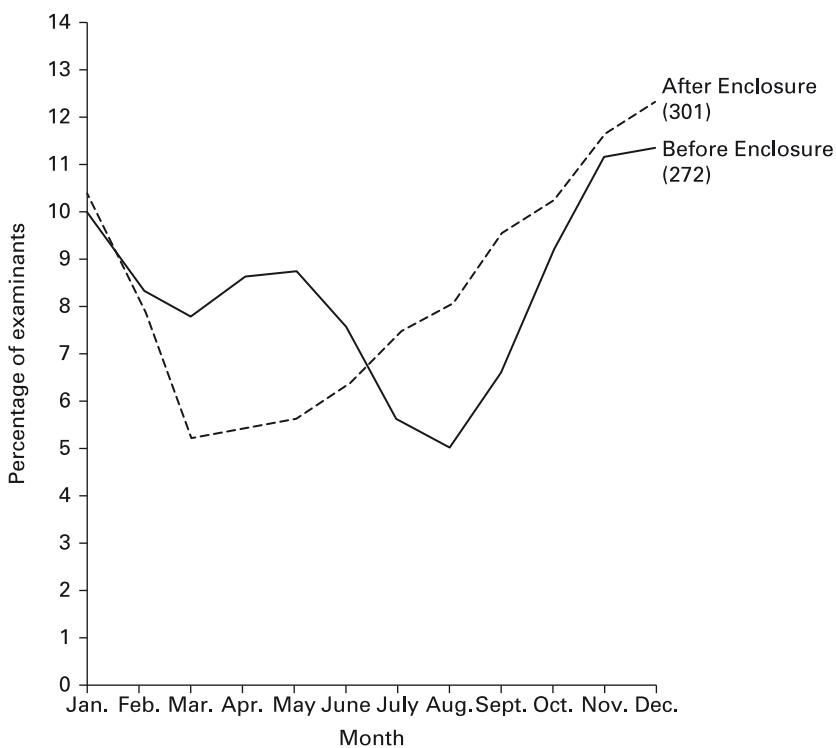


Figure 3.21 Distribution for female seasonal unemployment before and after enclosure in the counties of Bedfordshire, Cambridgeshire, Essex, Hertfordshire, Huntingdonshire, Norfolk and Suffolk. Unemployment is measured as a three-month moving average (see chapter 5).

Source: K. D. M. Snell, *Annals of the Labouring Poor* (Cambridge 1985), p. 156. For moving averages see pp. 129–31, 266.

Demographic analysis has been responsible for extending the use of graphical representations in further novel ways. Many examples can be found in Wrigley and Schofield's pioneering study of English population history published in 1981 and in their later volume based on family reconstitution evidence.⁸

Figure 3.22 is one example. It shows change in the 'dependency ratio' in England and Wales over time. The dependency ratio is calculated as the number of those aged 0–14 and 60 and over per 1,000 persons aged 15–59. The ratio can be seen to be climbing steeply as the birth rate began to rise, mortality overall stabilized and fell, and population therefore expanded prior to and during industrialization. Children aged 5–14 probably accounted for between 23 per cent and 25 per cent of the total population by the beginning of the nineteenth century compared with 6 per cent in 1961. Such a ratio helps to explain the prevalence of child labour in the late eighteenth and early nineteenth centuries and the pressure placed upon living standards of the many families with large numbers of children. The graph demonstrates this better than any other method of visual display.

A type of graph very commonly used in economics and in economic and social history is the **Lorenz curve**.

The Lorenz curve is a cumulative percentage curve that plots accumulated wealth (%) on the *y*-axis against cumulative population (%) on the *x*-axis.

The Lorenz curve gives an immediate impression of the level of inequality in wealth terms. It is used for comparisons between countries, or between different time periods,

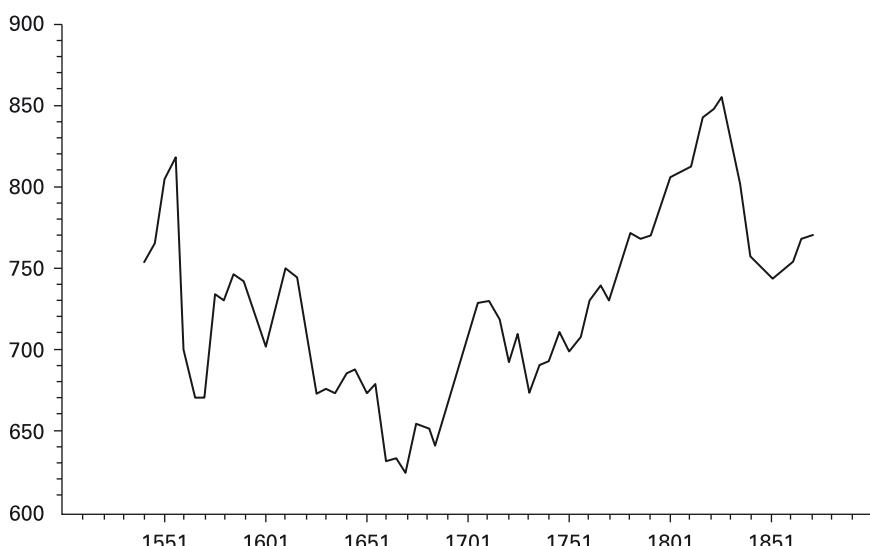


Figure 3.22 The dependency ratio, 1541–1871.

Note: The dependency ratio is taken as the number of those aged 0–14 and 60 or over per 1,000 persons aged 15–59.

Source: E. A. Wrigley and R. S. Schofield (1989) *The Population History of England, 1541–1587: A Reconstruction*, Cambridge: Cambridge University Press, p. 444.

History by Numbers

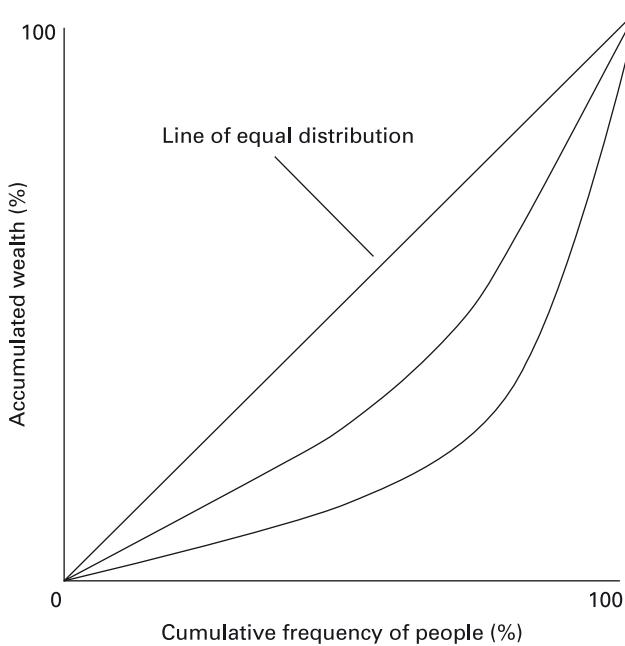


Figure 3.23 Lorenz curves (hypothetical data).

rather than as a quantitative measure of inequality. Figure 3.23 shows two Lorenz curves of different shapes indicating contrasting inequalities. The curve nearer to the straight line represents a society in which wealth distribution is more equal than is the case in the society represented by the Lorenz curve which is more distant from the line of equal distribution.

The skills of the historian must come to the fore in the interpretation of Lorenz curves and in criticizing the reliability and comparability of their sources. Inter-country comparisons are only useful if the sources upon which the respective Lorenz curves are based, are directly comparable. Recent controversial and inspiring work on historical shifts in wealth inequality over time in advanced capitalist economies has alerted us to some of these difficulties. It used to be assumed that wealth inequality would decline over time with the advancement of societies economically. This has now been challenged by using tax rather than expenditure data and by paying less attention to incomes and more attention to the impact of intergenerational transfers of fixed assets such as housing.⁹

Word clouds and similar figures in textual analysis

In recent decades there has been a huge growth in digitized textual sources available on the Internet for computer-aided statistical analysis. This has been accompanied by the development of software for textual research and lexical study of various kinds (further discussed in Chapter 8). Such research and study is largely based upon automated word

counts and the identification of word strings and phrases and their frequency. Software can also be used for finding patterns in text that might otherwise go unnoticed. The ability to analyse texts in this way, often on a scale that would not previously have been possible (without digitization and powerful computers), has been accompanied by various new devices for the display of the relevant results. One of the most common is the word cloud as shown in Figure 3.24. A word cloud is an image composed of words used in a particular text, or range of texts, where the size of each word indicates its frequency and hence its supposed importance.

Figure 3.24 shows the frequency of occurrence of certain words in *Domestic Duties* written by Frances Parkes and published in 1825.¹⁰ It is part of a larger study of female roles in local government and the public sphere in the late eighteenth and early nineteenth centuries, based upon a wide range of digitized (machine-readable) British local newspapers and domestic economy texts.¹¹ Word clouds of this type often use colour to differentiate the words. In the example shown here the prevalence of words such as 'governed', 'power', 'happiness' and 'mind' suggest that this advice book engages the reader with a world far beyond the domestic sphere. Whether it demonstrates the reality or strength of such connections is of course another matter and needs corroborative research, some of which is included in the article by Richardson.

It is worth noting that the textual sources used here were inaccessible to this sort of analysis before the possibilities opened up by mass digitization projects and the availability of many of their results on the Internet. Textual analysis of this kind is one of a range of new approaches to historical sources made possible only with progress of digitization and computer software that have occurred in recent decades. Such approaches



Figure 3.24 Word cloud analysis of Frances Parkes, *Domestic Duties* (1829).

Source: S. Richardson (2015) 'Petticoat politics in eighteenth and early nineteenth-century Britain: female citizenship revealed by the digital archive', Working Paper. Coventry, UK: University of Warwick (Unpublished), p. 17.

History by Numbers

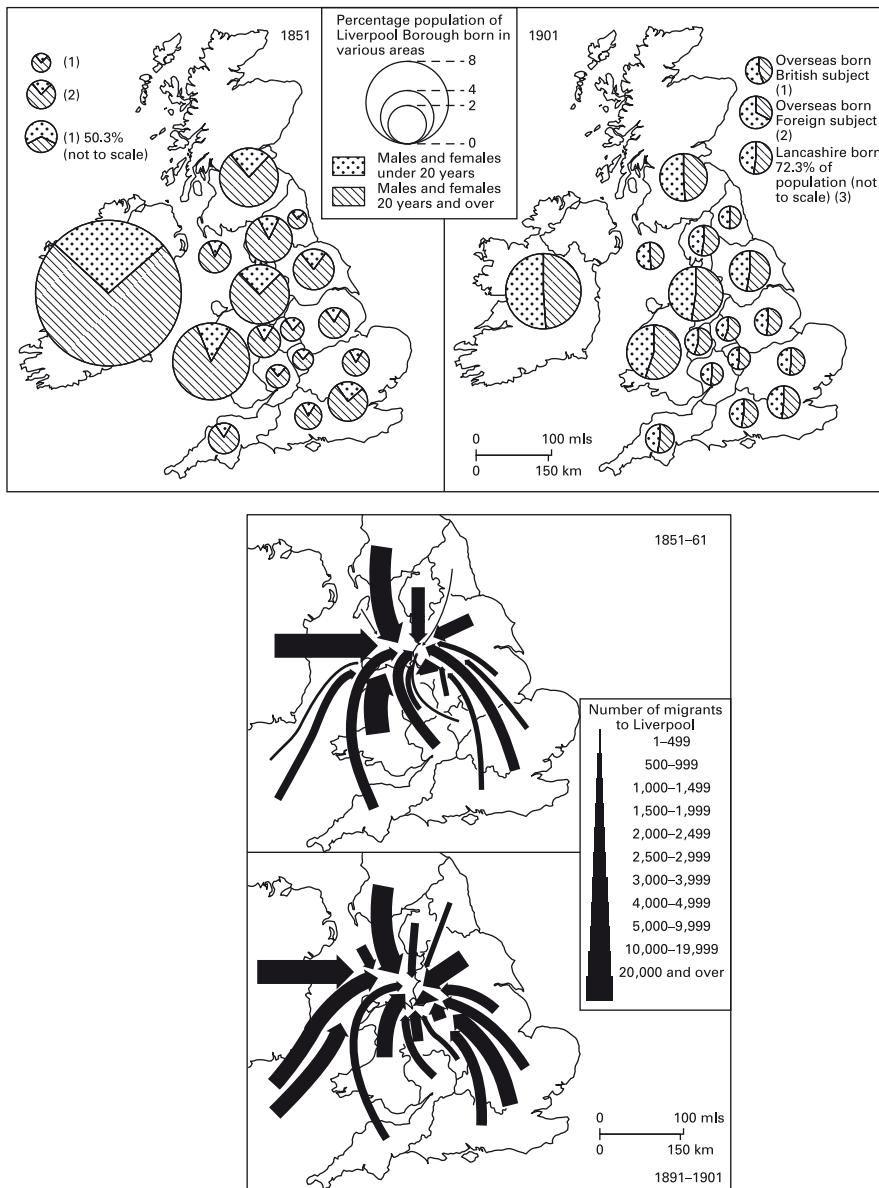


Figure 3.25 Cartograms showing age and origin of immigrants to Liverpool, 1851–1901.
 Source: R. J. Lawton, 'Population', in J. Langton and R. J. Morris (eds), *An Atlas of Industrialising Britain 1780–1914* (London 1986), p. 29.

are often facilitated by the fact that, as here, the digitized collections include comprehensive search tools that allow for such things as misspellings and synonyms to be taken into account (so-called 'fuzzy' searching). More advanced software enables common phrases and relationships between words to be identified and analysed: methods increasingly used in literary analysis and also very useful in historical work.

Cartograms

The role of **cartograms** (maps onto which graphs, symbols, pie charts and so on are superimposed to represent different variables) have grown increasingly popular in recent years especially with the expansion of computer software, notably Geographical Information Systems (GIS), which makes them much easier to prepare.

Figure 3.25 illustrates a historical use. It is drawn from *An Atlas of Industrialising Britain 1780–1914* (1986) which contains many such examples.¹²

Conclusion

This conclusion starts with a *warning*. The reordering, reclassification, regrouping and visual display of statistical information in charts, graphs or tables carries a heavy responsibility because it is easy to distort the appearance of series or runs of figures by choosing inappropriate display techniques, distorting classifications, or inappropriate scales of measurement (on the axes of a graph, for example). These can either exaggerate or underplay the character of the original data and create a misleading impression of the evidence. It must also be remembered that *all* rearranging and classifying of data loses something of the integrity and richness of the original source and this is something that should always be borne in mind. Even at this simple level of descriptive statistics, the important first step is always to look closely at the figures and their origin and to consider, as a historian, the likely omissions, biases and distortions of the evidence. Such problems need to be thought about in relation to the purpose for which the data will be employed, that is, the research questions being asked. Careful judgements at this point are important in determining whether further manipulation or analysis is wise or worth doing.

Having sounded the alarms, one can also say that elementary descriptive statistics, when employed with care and attention, can be a very useful tool for the initial summary and display of raw data. Manipulation into a matrix, frequency distribution, chart, graph or figure can go a long way in clearly highlighting the characteristics of the data and in helping to pose further analytical questions.

Further reading

- Daly, F., D. J. Hand, M. C. Jones, A. D. Lunn and K. J. McConway, *Elements of Statistics* (Harlow 1995), Chapter 1.
- Feinstein, Charles and Mark Thomas, *Making History Count: A Primer in Quantitative Methods for Historians* (Cambridge 2009).
- Foster, Liam, Ian Diamond and Julie Jeffries, *Beginning Statistics: An Introduction for Social Scientists* (2nd edition, London 2015) Chapters 1–3.
- Gonick, Larry and Woolcott Smith, *The Cartoon Guide to Statistics* (New York 1993).
- Hanagan, T., *Mastering Statistics* (3rd edition, London 1997), Chapters 4 and 5.

History by Numbers

- Haskins, Loren and Kirk Jeffreys, *Understanding Quantitative History* (Cambridge, MA 1991),
Chapters 1–2.
- Jarausch, Konrad H. and Kenneth A. Hardy, *Quantitative Methods for Historians: A Guide to
Research, Data, and Statistics* (Chapel Hill 1991).
- Tufte, E., *Visual Explanations: Images and Quantities, Evidence and Narrative* (Cheshire, CT
1997).

CHAPTER 4

SUMMARIZING DATA: AVERAGES AND DISTRIBUTIONS

This chapter considers the nature of datasets and the vectors within them. It suggests simple ways in which distributions of values can be described, summarized and analysed. A **distribution**, as we learned in Chapter 3, is a range of values observed for any one variable. Most column vectors in datasets consist of a distribution of values.

MEASURES OF CENTRAL TENDENCY

One of the first things that one may wish to do with a distribution of values is to calculate the **average** value. The average is an important summary characteristic but, as we shall see, averages must be chosen with care.

An **average** provides a value around which a set of data is located. It is a measure of **central tendency** in the data. Calculating the central tendency in interval or ratio data is often the first stage of an investigation. There are three commonly used measures of average:

The **arithmetic mean** usually referred to as just the **mean**;

The **median**;

The **mode**.

Each of these measures the average or central tendency of a distribution in a different way.

The choice of measure depends upon the nature of the distribution and the purpose for which the average is being calculated.

The mean

The **mean** is the average as it is most commonly understood and calculated: formed by adding all the values together and dividing by the number of observations. It is used only for interval data. The advantages of the mean are that:

- (a) It takes account of all of the values;
- (b) There are measures of dispersion that can be used with it (see below).

History by Numbers

The major disadvantage is that it is sensitive to untypical extreme values: the value of the mean may be badly distorted away from the typical experience by the presence of one or two unusually large or small outlying values.

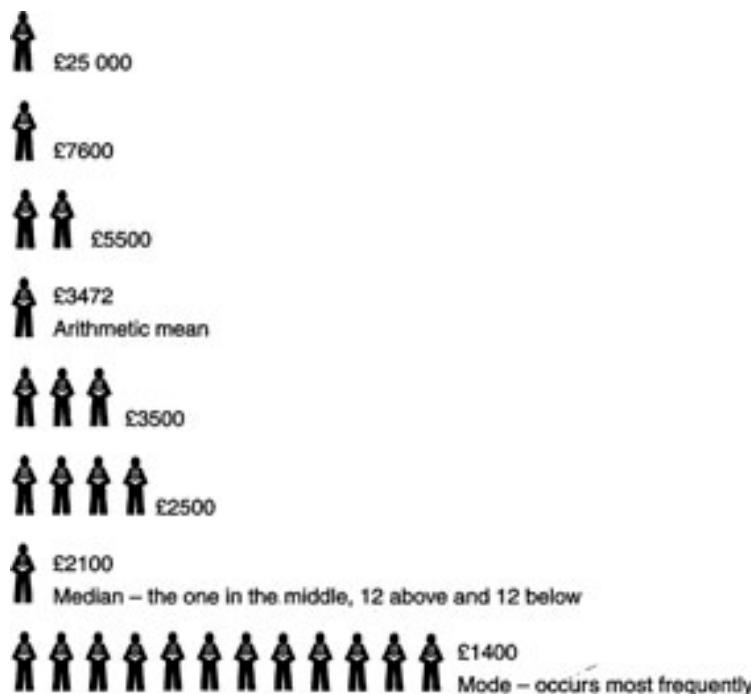


Figure 4.1 Pictogram of white-collar salaries in a firm in the 1950s.

Source: Based on Darrel Huff, *How to Lie with Statistics* (London 1973), p. 33.

In Figure 4.1, a pictogram showing average incomes in a business firm in the early 1950s, the mean would be a poor indicator of average experience because its value is inflated by the income of one man (the boss?) at the pinnacle of the earnings pyramid. The **mean** is usually represented by the symbol \bar{X} and the formula for calculating the mean is given as:

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Where

\bar{X} is the mean of vector X ;

X_i is the value of the variable for case i ;

N is the number of observations;

Σ is 'the sum of'.

The mean of land tax payments in Table 3.17, for example, can be calculated as:

$$\text{mean} = \frac{\text{sum of all payments}}{\text{number of tax payers}} = \frac{\text{£202 10s. 0d.}}{47} = \text{£4 6s. 1d.}$$

It should be noted that in this case, the mean is again not a very good measure of the average or typical payment because of the existence of one or two atypically large payers. (Atypically large or small values in a distribution are usually referred to as **outliers**.)

The mean can also be calculated from a frequency distribution using the formula:

$$\bar{X} = \frac{\sum_{i=1}^k f_i X_i}{N}$$

Where

X_i is value of the variable for group i ;

f_i is the frequency with which those values occur;

k is the number of groups;

N is the number of cases from which the frequency distribution has been compiled.

Thus, if we take the frequency distribution in Table 3.11, drawn from Table 3.10, we can calculate the mean prison sentence as:

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^6 f_i X_i}{N} \\ &= \frac{[(3 \times 7) + (2 \times 10) + (1 \times 12) + (5 \times 14) + (8 \times 15) + (2 \times 20)]}{21} \\ &= \frac{(21 + 20 + 12 + 70 + 120 + 40)}{21} \\ &= \frac{283}{21} \\ &= 13.5\end{aligned}$$

Where

f_i in this case is the number of prisoners;

X_i is the sentence length;

N is the total number of prisoners;

k is the number of different sentence lengths represented.

Yet again the mean is not really the best measure of the average for this distribution because no prisoner is serving a 13.5-year sentence. All sentences are in whole years but more importantly there is an obvious candidate for the most typical experience which is 15 (this is by far the most commonly occurring experience, known as the mode; see below).

The mean provides the most justified measure of average when a distribution has few outliers that are likely to distort the mean, and when the values of the variable seem to be fairly evenly spread around a central value. Many measures occurring in nature,

History by Numbers

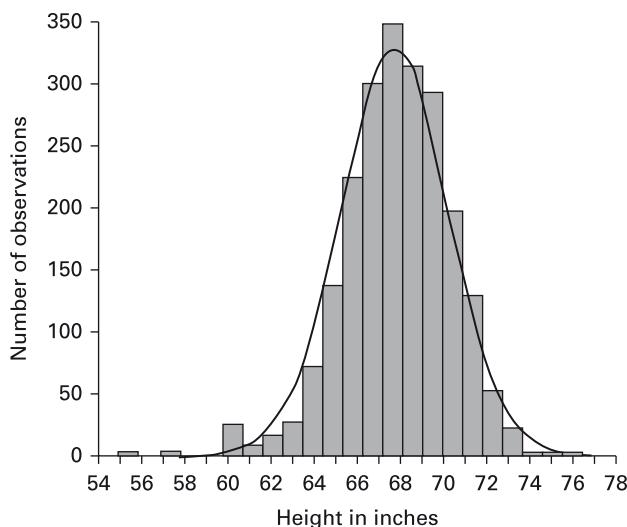


Figure 4.2 Height distribution of US passport applicants, 1830–1857.

Source: John Komlos, ‘On the nature of the Malthusian threat in the eighteenth century’, *Economic History Review*, 52, 4 (1999), p. 736. In this figure the histogram is compared for analytical purposes with the shape of the normal distribution (see pp. 112–113).

such as human physical attributes, other biological data and observations of social or economic characteristics such as the sizes of households, values of industrial firms, numerically scaled educational levels, social skills or wages in a population, tend to cluster evenly (above and below) an average measure, with very few observations lying outside of a certain range. This type of finite and symmetrical distribution (known as a binomial or bell-shaped distribution) is a common one in historical evidence (although the variables that historians are interested in are often distributed in other ways too).

Figure 4.2 provides an example of a bell-shaped distribution. There is a fairly even spread of observations above and below the mean, tapering off symmetrically. For further discussion of the properties of bell-shaped distributions, see p. 112.

The median

The **median** is the observation that lies at the centre or middle of a distribution when all of the observations or values are ranked in size order. It can be used with ordinal or interval data. When there is an even number of cases the median is the mean of the two middle ranking values. The advantages of using the median are:

- (a) it is immune from the influence of extreme values
- (b) it has some measures of dispersion associated with it (see below pp. 101–104).

The median is a better way of calculating the average level of tax paid from Table 3.17, than was the calculation of the arithmetic mean, above, because it is immune from the influence of the two largest and untypical tax payers (outliers) (Sir Wats Horton,

Gentleman, who held 24 separate parcels of land and George Stansfield, merchant, who owned 15). The median tax payment is £1 0s. 6d. compared with the mean of £4 6s. 1d.

It is possible to *estimate* the median of a grouped frequency distribution which is useful if the original figures are not available. This is done by assuming that the values of items in the class containing the median are distributed evenly, that is that the median falls in the middle of that class. The probability of this occurring increases with the size of the dataset so this may be a decisive consideration in adopting this calculation. The median value of land tax paid in Sowerby in 1782 can be calculated roughly from Table 3.18 as £3, which may be a useful enough approximation depending upon the nature of the enquiry, but the size of the dataset might warn against this method.

The mode

The **mode** is the most commonly occurring observation. It can be used with nominal, ordinal or interval data and is the only average one can use with nominal data. The advantage of the mode is that it represents the most common experience or occurrence but the disadvantage is that it takes no account of other observations, has no measure of dispersal associated with it and can be entirely misleading if there is more than one commonly occurring observation (as in a bi-modal or tri-modal distribution of which more below). It is a useful measure when a distribution is not spread evenly around a central value but is of limited use when data are very dispersed.

In our example above concerning sentence lengths of prisoners, the mode is a better way of expressing the average prison sentence than the mean as the distribution is not widely spread and there is a very obvious common experience of 15 years.

In a grouped frequency distribution the **modal class** is the one with the highest frequency. In Table 3.18, for example, the modal class of land tax payers is 5s.–£1 which contains 32 per cent of the observations.

In our pictogram in Figure 4.1 the mode (£1,400) and the median (£2,100) are both better expressions of average than the mean (£3,472) because they are less affected by the one high outlier.

The geometric mean

The **geometric mean** is another average but it is less commonly used than the mean, median or mode. The geometric mean is defined as the N th root of the product of the distribution (where N is the number of items in the distribution). The geometric mean is used only with interval data, mostly in averaging growth rates or indices of growth. (An index-plural indices-is a series expressed in percentage terms as explained in Chapter 5.)

To calculate the geometric mean one multiplies all the N values of a variable, X , together and then one takes the N th root:

$$\text{geometric mean} = \sqrt[N]{X_1 X_2 X_3 \dots X_N}$$

History by Numbers

This may also be written as:

$$\text{geometric mean} = (X_1 X_2 X_3 \dots X_N)^{1/N}$$

Note that there is no need for multiplication signs: $X_1 X_2$ is the same as. $X_1 \times X_2$.

Example

If the price of commodity A rises from £25 to £50 this is an increase of 100 per cent

If the price of commodity B rises from £80 to £100 this is an increase of 25 per cent

The mean increase of the two price rises is $\frac{125 \text{ per cent}}{2} = 62.5 \text{ per cent}$

The geometric mean = $\sqrt{100 \times 25}$ per cent=50 per cent

Which of the two measures of average growth detailed above one should use is open to debate and depends upon the researcher's purpose. The geometric mean gives less weight to extreme values than the arithmetic mean but there is no measure of dispersion associated with it. Growth rates can also be measured and expressed in other ways (see Chapter 5) and these methods are often preferred to either the arithmetic or the geometric mean.

Choice of average

It is not always easy to make a clear-cut decision about which measure of average (mean, median or mode), is the best reflection of typical experience given the character of the data. Sometimes it will depend upon the questions that one is asking about the evidence. The mode will be favoured where the most common experience is desired, the median where better knowledge of the impact of the distribution on average experience is needed, and the mean will be chosen where account must be taken of all of the observations equally (best done where there are few or no outliers and where the distribution is not markedly skewed – see below) Often two or all three of the measures are stated together. The differences between them provide a good indication of the nature of the distribution of values.

For example, in Table 4.1 taken from E. A. Wrigley's study of marriage ages in early modern Colyton, Devon, all three of the measures of average marriage age are given. This is because each highlights different features of the data upon which Wrigley comments and each gives useful additional information. Where the three measures of average are given together in this way an indication of the shape of the distribution as a whole can be visualized. The data on male and female average marriage ages in Table 4.1 can be seen immediately to be 'skewed' because for both men and women and for all of the time periods the mean is greater than the median which in turn is greater than the mode. This is characteristic of a positively skewed distribution (see below Figure 4.10a). The use of the

Table 4.1 Age at first marriage in Colyton, 1560–1837

	<i>Number</i>	<i>Mean</i>	<i>Median</i>	<i>Mode^a</i>
Men				
1560–1646	258	27.2	25.8	23.0
1647–1719	109	27.7	26.4	23.8
1720–1769	90	25.7	25.1	23.9
1770–1837	219	26.5	25.8	24.4
Women				
1560–1646	371	27.0	25.9	23.7
1647–1719	136	29.6	27.5	23.3
1720–1769	104	26.8	25.7	23.5
1770–1837	275	25.1	24.0	21.8

^a The mode is interpolated from the mean and the median and not derived directly from the data.

Source: E. A. Wrigley, 'Family limitation in pre-industrial England', *Economic History Review*, 19, 1 (1966), p. 86.

mean with the median is a common way of roughly indicating the shape of a distribution of values.

Another example of a positively skewed distribution where the mean is biased upwards, away from average experience by the presence of a relatively small number of very high values, is the distribution of wealth in most societies. Recent controversial research has suggested that inequality in terms of accumulated wealth, as well as in terms of income distribution has been growing in developed countries in the last few decades.¹ Around the turn of the millennium, the Institute of Fiscal Studies reported that although average wealth was growing in Britain, the distribution was becoming more unequal. Mean wealth (in accumulated savings and assets) of £7,136 contrasted markedly with a median of only £750! In addition, 30 per cent of the population had no savings outside of their home and pension and around 10 per cent (mostly single parents and out-of-work couples) had no savings at all.² The skew of the distribution both of income and of wealth is much more extreme in the United States largely because the redistributive impact of state spending, social security and healthcare spending is much more muted. The top 1 per cent of society in the USA in 2012 owned 40 per cent of the wealth (compared with 23 per cent in 1978). The top 0.1 per cent held 22 per cent (compared with 7 per cent in 1978).³ In Figure 4.3 it can clearly be seen that wealth holding is skewed: the lowest half of the population own only 9 per cent of wealth, the top 20 per cent own 65 per cent (2008–2010 figures). (The calculation and use of percentiles and deciles is covered later in this chapter.)

Another example to illustrate the nature of skew in distributions is provided by Botticini's analysis of the marriage market in fifteenth-century Tuscany which is included as an exercise later in this volume.⁴ The median and the mean are displayed together for a number of different variables as shown in Table 4.2. The means are higher than the

History by Numbers

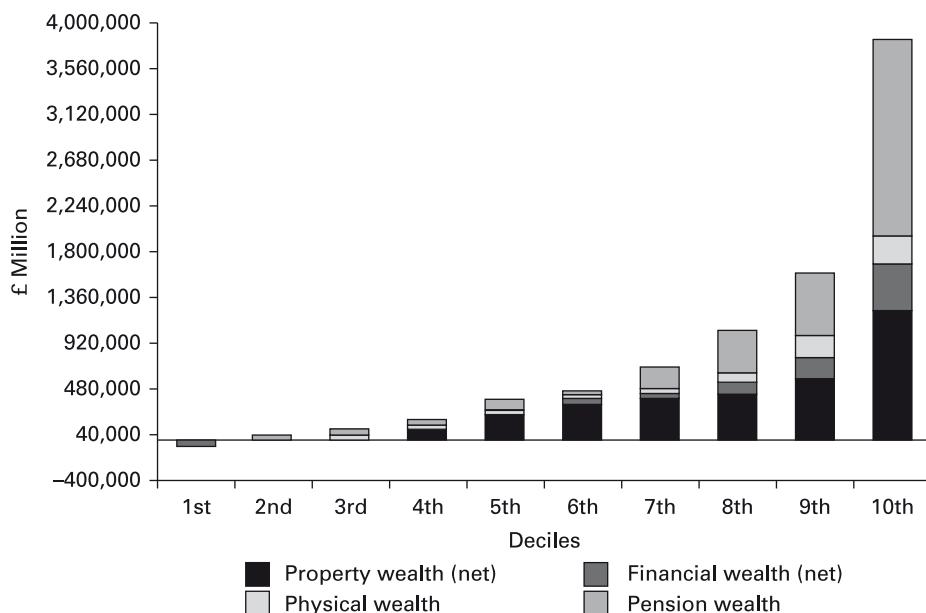


Figure 4.3 Wealth distribution in England.

Source: Office of National Statistics, Wealth and Assets Survey 2008–2010.

Table 4.2 Summary statistics of marriages in Cortona, 1415–1436

	Mean	Median	Standard deviation
Dowry (florins)	125.5	70	105.9
Groom's age (years)	28.1	27	8.3
Bride's age (years)	18.8	18	4.7
Groom household's wealth (florins)	609.7	164	1692.84
Bride household's wealth (florins)	700.7	196	1997.66
Number of children in grooms' households	2.25	2	1.87
Percentage of daughters in grooms' households	0.08	0	0.17
Number of children in brides' households	3.14	3	2.33
Percentage of daughters in brides' households	0.65	0.6	0.27
N		224	

Note: the marriages refer to households living in the town of Cortona and in 44 villages in its countryside.

Source: M. Botticini, 'A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town, 1415–1436', *Journal of Economic History*, 59, 1 (1999), p. 108.

medians for all of the variables again indicating the presence of positively skewed distributions. The standard deviation, as used here, is a measure of dispersion of the data around the average and is explained in the next section.

MEASURES OF DISPERSION

An average on its own tells us very little about the entire population: in particular, it says nothing about how divergent from the average is the distribution of individual observations. All distributions are not only clustered around central points but also spread out, or dispersed, around them. The **range** is a first indication of dispersal. The range of a set of data is literally its spread: the highest value of the distribution minus the lowest. The range is often used with the mode but can be used with any interval data.

There are a number of more sophisticated measures of dispersion which can be used with the mean and the median.

Dispersion around the mean: standard deviation and variance

Many very different distributions can have the same mean. For example, all three of the distributions in Table 4.3 have a mean of 45.87 despite the fact that A is widely dispersed (range 99) whilst C is closely clustered (range 4) and B is influenced by the presence of one extreme atypical value (an ‘outlier’) and has the largest range because of this. Hypothetical data are used here to make the differences between the three distributions very clear.

Table 4.3 Distribution of defamation cases in three English courts, 1680–1687

Year	<i>Distribution</i>		
	A	B	C
1680	100	20	48
1681	88	28	47
1682	70	22	46
1683	50	45	45
1684	30	16	45
1685	20	167	45
1686	8	40	44
1687	1	29	47
Total	367	367	367

Source: Hypothetical data.

History by Numbers

Because the average on its own tells us little about the entire population it is almost always used with some indication of the spread of data. A measure of dispersion tells us to what extent the values of a distribution are, or are not, bunched around the average. The measures of dispersion most commonly used with the mean are the **variance** and the **standard deviation**.

The **variance** is the average of the squares of the deviations from the mean. It is calculated by adding the square of the deviations of the individual values from the mean of the distribution together and dividing this sum by the number of items in the distribution. The following formula achieves this:

$$\text{variance} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

Where

\bar{X} is the mean;

X_i is the value of the variable for row i ;

N is the number of observations.

The **standard deviation** is another measure of dispersion around the mean. It is normally represented by the letter s or by the abbreviation SD. It is found by applying the formula for the variance and then taking the square root. The variables, and \bar{X} , X_i and N are as defined already. The variance is always equal to the square of the standard deviation (that is, s^2).

$$s = \sqrt{\left(\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N} \right)}$$

In distribution A, $s = 34.77$ and the variance is 1209.11.

In distribution B, $s = 46.69$ and the variance is 2180.36.

In distribution C, $s = 1.27$ and the variance is 1.61.

See Table 4.4 for a partial breakdown of the calculations.

The greater the dispersion, the larger the standard deviation and the variance. In each case s is expressed in the original units, in this example in court cases.

The standard deviation can also be directly calculated from a grouped frequency distribution by applying the formula:

$$SD = \sqrt{\frac{\sum f D\bar{x}^2}{\sum f} - \left(\frac{\sum f D\bar{x}}{\sum f} \right)^2} \times \text{class interval}$$

Here f is the frequency and $D\bar{x}$ is the deviations from the mean (or the assumed mean). Statistical software applied to data in a matrix or an electronic spreadsheet makes this calculation very straightforward.

The formula for the standard deviation (s) takes into account the amount that each value deviates from the mean (the $X_i - \bar{X}$ part of the formula), which is what makes it so much more useful, in most cases, than the range.

Table 4.4 Statistics relating to Table 4.3

Year	i	Distribution					
		A		B		C	
		X_i	$X_i - \bar{X}$	X_i	$X_i - \bar{X}$	X	$X_i - \bar{X}$
1680	1	100	54.1	20	-25.9	48	2.1
1681	2	88	42.1	28	-17.9	47	1.1
1682	3	70	24.1	22	-23.9	46	0.1
1683	4	50	4.1	45	-0.9	45	-0.9
1684	5	30	-15.9	16	-29.9	45	-0.9
1685	6	20	-25.9	167	121.1	45	-0.9
1686	7	8	-37.9	40	-5.9	44	-1.9
1687	8	1	-44.9	29	-16.9	47	1.1
$\sum_{i=1}^N (X_i - \bar{X})^2$		9673		17442		12.88	
Variance		1209.12		2180.36		1.61	
Standard deviation		34.77		46.69		1.27	

Note: For all distributions, the number of observations, N , is 8, the number of court cases is 367, and the average, \bar{X} is 45.9. Note also that the square of a negative number is positive (i.e. -15.9 squared = $-15.9 \times -15.9 = +252.81$).

Source: Hypothetical data.

The Z score

Use of a measure called the **Z score** is a common method in the social science and historical literatures. A Z score is the number of standard deviations which an observation is above the mean (if it is positive) or below the mean (if it is negative). Where Z scores are used the standard deviation becomes a sort of yardstick for comparative purposes. Distributions of Z scores can be created that enable the dispersion of different distributions to be compared. The standard deviation itself is no good for this because it is expressed in the original units of measure, for example, dollars, pounds sterling, persons, cows. Z scores provide a universal unit for measuring dispersion. Because Z scores have standard values they are sometimes called standard scores.

History by Numbers

In research on growth stunting in children born in Rwanda between 1987 and 1991 the relative impact of crop failures and the military conflict was the focus of attention in a study employing Z scores. It was found that in poor and non-poor households, boys and girls born during the Rwandan conflict, in regions experiencing fighting, were negatively affected, with height-for-age Z scores 1.05 standard deviations lower than the norm. Conversely, only girls were negatively affected by crop failure, with girls exhibiting 0.86 standard deviations lower height-for-age Z scores, the impact being worse for girls in poor households. This suggested that girls bore the brunt of dietary restriction in times of crop failure but that both sexes were similarly affected by conflict.⁵ In a study of height and living standards in China between 1979 and 1995 (used as an exercise later in this volume), Chinese heights were compared to international reference standards of Z scores for such distributions at various ages.⁶

Dispersion around the mean: the coefficient of variation

The **coefficient of variation** is another measure of the extent to which a variable differs from its mean. It is simply the standard deviation (s), divided by the mean and is generally expressed as a percentage:

$$\text{coefficient of variation} = \frac{s}{\bar{X}} \times 100 \text{ per cent}$$

Because it is expressed as a percentage it can be used to compare the dispersion of distributions of different sorts of variables one with another. The coefficient of variation is normally only calculated for this purpose – to compare the degree to which two variables differ from their respective means. It is not possible to use standard deviations for this because standard deviations are expressed in the original units of the variable, for example, persons, exports, strikes, ploughs, hearths, looms and so on, whereas the coefficient of variation is always a percentage.

If we were told that the three distributions in Table 4.3 were not all court cases but that each distribution related to a different variable we would need, for comparative purposes, to calculate the coefficient of variation. For example, if the dataset described the assets of eight farmers in the early nineteenth century with:

- series A the number of cows
- series B the value of seed on hand in £
- series C the value of land and farm buildings in £thousands

We might wish to calculate the coefficient of variation to see the extent to which the different sorts of assets of these farmers varied from the average experience. The coefficients of variation of the three distributions are:

Distribution A: 0.76 per cent

Distribution B: 1.02 per cent

Distribution C: 0.03 per cent

The coefficient of variation is also used in comparing the variation of certain measures at different time periods or for different countries because standard measures are far easier to compare than original units. Tables 4.5 and 4.6 are drawn from an article by Jeffrey G. Williamson entitled 'Globalisation, convergence and history'. His estimates of coefficients of variation of real wages, 1854–1939 and of coefficients of variation of Gross Domestic Product (GDP) per worker hour, for the OECD, 1870–1938, support his argument that growth convergence is linked to globalization and that convergence was arrested in the period 1914–1950.⁷

Another example is provided by Lazerev and Gregory's study of vehicle allocation in the Soviet command economy of the 1930s (included as an exercise later in this volume). They show 'satisfaction rates' (the ratio of allocated to requested vehicles). They show that the Soviet system began with an enormous excess demand for vehicles but was able to satisfy nine out of ten consumers by early 1937. Especially during supply shocks, such as that in 1932, however, the Dictatorship favoured 'preferred customers' such that there was a very uneven 'actual' distribution of satisfaction shown by a coefficient of variation that reached 83.5 per cent. Some figures from the research are given in Table 4.7. (Rows 3 and 4 can be ignored for the moment. They are explained in Chapter 6.)

Table 4.5 Coefficients of variation of real wages, 1854–1939

Year	Full sample ^a			Full sample less North America ^b			Full sample less North America and Iberia ^c	
	C(13)	C(17)	C(16)	C(12)	C(15)	C(14)	C(10)	C(13)
1854	0.326			0.308			0.340	
1870	0.254	0.255		0.224	0.223		0.229	0.232
1890		0.199			0.114			0.102
1913		0.191			0.068			0.039
1914			0.103			0.085		0.068
1926			0.148			0.146		0.138
1927		0.188	0.147		0.186	0.142		0.131
1939		0.285			0.200			0.138

^a The 'full sample' included the following 13 countries until 1870: Australia, the United States, Belgium, France, Germany, Great Britain, Ireland, Netherlands, Norway, Spain, Sweden, Brazil and Portugal; in 1870 the following four countries were added to the sample: Argentina, Canada, Denmark and Italy; Portugal dropped from the sample from 1914 to 1926 and then rejoined.

^b 'Full sample less North America' excludes Canada and the United States, implying that we start with 12 countries and then increase to 15 in 1870; again, Portugal dropped from the sample between 1914 and 1926.

^c 'Full sample less North America and Iberia' excludes the United States, Canada, Spain and Portugal, implying that we start with 10 countries and expand to 13 in 1870.

Note: The number of countries in the sample, x , is indicated by the column heading $C(x)$.

Source: J. G. Williamson, 'Globalisation, convergence and history', *Journal of Economic History*, 56, 2 (1996), p. 280.

History by Numbers

Table 4.6 Coefficients of variation of gross domestic product (GDP) per worker-hour, 1870–1938

Year	<i>Full sample^a</i>		<i>Full sample less North America^b</i>	
	C(15)	C(13)	C(13)	C(13)
1870	0.153		0.169	
1890	0.118		0.122	
1913	0.107		0.088	
1929	0.110		0.080	
1938	0.090		0.054	

^a The ‘full sample’ includes Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom, and the United States; it does not include Japan.

^b The ‘full sample less North America’ drops Canada and the United States from the full sample.

Source: J. G. Williamson, ‘Globalisation, convergence and history’, *Journal of Economic History*, 56, 2 (1996), p. 280.

Table 4.7 Vehicle allocation in Soviet Russia during supply shocks and periods of normality

	1932		1934		1934		1937	1937
	Planned	Actual	3rd quarter		Planned	Actual	2nd quarter	4th quarter
			Planned	Actual				
(1) Average satisfaction rate, %	31.4	15.0	41.7	39.6	53.6	51.2	87.6	46.5
(2) Coefficient of variation in satisfaction rates, %	26.4	83.5	47.8	48.5	37.5	37.8	18.3	53.3
(3) Pearson R ² correlation	0.996	0.919	0.962	0.962	0.807	0.778	0.98	0.747
(4) Spearman rank correlation	0.976	(0.542)	0.957	0.951	0.861	0.881	0.978	0.645

Source: Valery Lazerev and Paul R. Gregory, ‘The wheels of the command economy: allocating Soviet vehicles’, *Economic History Review*, 55 (2), (2002), pp. 324–348, p. 333.

In the study of living standards and stature in China, mentioned above, the coefficient of variation was used to show that variation in heights amongst rural children was greater than amongst urban children and that this increased over time.

Rank order dispersal measures

These are commonly used with the median. The median is only one of a range of measures that summarize data according to their rank order. The median divides the ranked distribution into half. The first **quartile** (Q_1) is defined as the middle number between the smallest number and the median of the dataset. The second **quartile** (Q_2) is the median of the data. The third **quartile** (Q_3) is the middle value between the median and the highest value of the dataset. Others measures commonly used in the same way are:

quintiles

deciles

percentiles

The three **quartiles** divide the ranked distribution into 4 equal parts.

The four **quintiles** divide the ranked distribution into 5 equal parts.

The nine **deciles** divide the ranked distribution into 10 equal parts.

The 99 **percentiles** divide the ranked distribution into a hundred equal parts.

Consider the distribution of 20 observations in Table 4.8, ranked in size order, which derives from an archaeological project.

As mentioned above, the **median** can also be expressed as the second quartile (Q_2) and the measure of dispersion often used with the median is the **interquartile range**. This is the difference between the first and the third quartiles (Q_1 and Q_3). In the example in Table 4.8, the interquartile range is:

$$Q_3 - Q_1$$

$$= 10.5 - 4.5$$

$$= 6$$

Sometimes this is divided by two to form the **semi-interquartile range** or **quartile deviation** which would in this example be 3.

The ninth decile of the distribution in Table 4.8 is 11.5. This distribution is too small to have percentiles. Percentiles can be calculated only where there are at least a hundred observations.

Figure 4.4 is a cartogram demonstrating the inequality of Russian landholding prior to the Revolution. The distributions of 'Private holdings' and of 'All types of property' are divided into quartile ranges which is a level of detail sufficient to make the main point:

History by Numbers

Table 4.8 Number of archaeological remains found by twenty postgraduate assistants

<i>Number of hearths:</i>	<i>Deciles</i>	<i>Quintiles</i>	<i>Quartiles</i>
			<i>Q</i>
2			
2			
3	1st = 2.5		
4			
4	2nd = 4.0	1st = 4.0	
5			1st = 4.5
5	3rd = 5.5		
6			
6	4th = 6.0	2nd = 6.0	
7			
7	5th = 7.5		2nd (median) = 7.5
8			
8	6th = 8.5	3rd = 8.5	
9			
9	7th = 9.5		
10			3rd = 10.5
11			
11	8th = 10.5	4th = 10.5	
11			
12	9th = 11.5		
23			

Note: Spreadsheet functions give slightly different results from the above first approximations as they take account of any skew in the distribution. In the example above the first quartile is really 4.75. The third quartile is 10.25 and the ninth decile is 11.10. The interquartile range is 5.5.

Source: Hypothetical data.

that there were clear local concentrations of wealth and a marked variation in average wealth holding in different regions. The **Gini coefficient** is a measure of statistical dispersion intended to represent the income distribution of a nation's residents, and is the most commonly used measure of inequality. A Gini coefficient of 0 denotes a completely equal distribution and of 1 (10, or 100 depending upon choice of scale) denotes complete inequality (where one individual earns all the income). Further discussion of Gini coefficients is included in Example 3 below. In our Russian example, property was more unequally distributed around St Petersburg and Odessa, and in the north-east of the country, on the eve of the Revolution, which may have created particular tensions in those areas. As the article also shows, income inequality closely matched the geographical distribution of land- and property-holding inequality.

Summarizing Data: Averages and Distributions

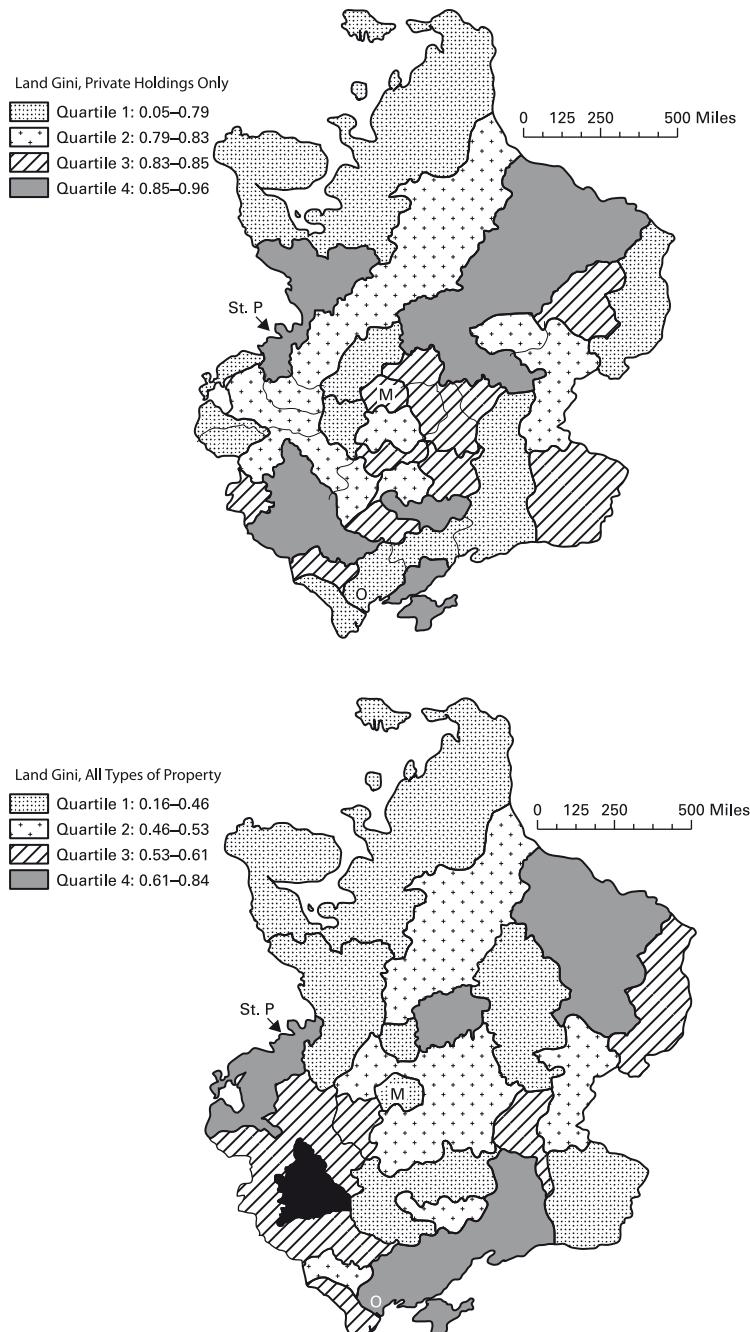


Figure 4.4 The geography of landholding inequality in Russia c. 1905.

Source: Peter H. Lindert and Steven Nafziger, 'Russian inequality on the eve of Revolution', *Journal of Economic History*, 74 (3), (2014), pp. 767–798, p. 780.

History by Numbers

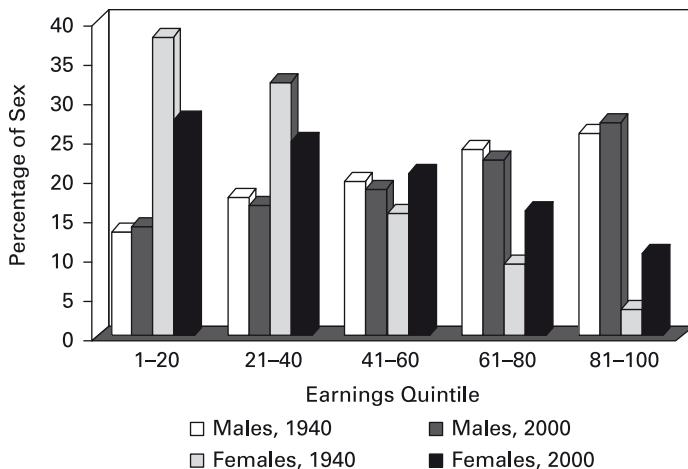


Figure 4.5 Real personal earnings quintiles for non-farm year-round workers by sex, 1940 and 2000.
Source: M.B. Katz, M.J. Stern and J.J. Fader, 'Women and the paradox of economic inequality in the twentieth-century', *Journal of Social History*, 39 (1), (2005), p. 78. © Oxford University Press.

Quartiles and quintiles are often used as divisions in bar charts as in Figure 4.5. This shows the earnings of non-farm workers by sex in the USA, 1940–2000. The authors use employment data from the census to explore the ‘paradox’ of inequality in twentieth-century America, that is, the enduring coexistence of inequality with individual and group social mobility. Figure 4.5 shows that women’s earnings were represented in a much wider range of the income scale by 2000 than ever before. The use of quintiles is effective here: the income of women working full-time is expressed in earning quintiles by gender for 1940 and 2000. The pattern immediately visible in Figure 4.5 is the persisting income disparity between men and women both in 1940 and 2000. Women were more likely to be found at the bottom quintile and much less likely to be at the top quintile of earnings. In 2000, the gender disparities are much less sharp than in 1940 but just as apparent. In their wider analysis the authors demonstrate that women had entered into a much wider variety of occupations and industries by 2000 than they had half a century earlier but that significant inequality of earnings emerged within the female workforce, as they had within the male labour force. Higher education played a key role in emerging income differences.

An advantage of the **interquartile range** and the **quartile deviation** is that they are immune from the influence of very small or very large values. This can be an advantage if there are just a few extreme outliers that would seriously effect alternative measures of dispersal such as the standard deviation.

More examples of analysis of distributions from history

If we are researching a historical question that hinges upon the nature of a dataset, choosing and applying the most appropriate measure of central tendency and dispersal

are likely to be crucial to the arguments which may be made. We now consider four examples from historical research of such choices and applications discussing their advantages and any disadvantages or weaknesses.

Example 1: Eighteenth-century slopsellers

Beverly Lemire's analysis of male and female activity as slopsellers in London in the later eighteenth century provides our first very simple example. Slopsellers were suppliers of clothing to the Navy. The types of work clothes supplied included knitted caps, stockings, shirts, waistcoats, shoes, handkerchiefs, drawers, and 'blew' suits. Slopsellers had to bid for the contracts and once secured the clothing would be delivered to the ship in question where it was stored in slop chests under the care of the purser. During a voyage and as necessary, the purser would sell the garments to the crew keeping a tally for the slopseller.

Comparing the distribution by insured wealth of male and female slopsellers gives an indication of the difficulties that women faced in running businesses of this kind in the period. There were much fewer women running these businesses than men and although both distributions are skewed in favour of smaller sized concerns the female-headed business distribution is more skewed than the male. Social attitudes to women working at all levels but particularly as business owners, together with legal obstacles concerning the restriction upon women owning property in their own right and acting as an independent legal entity, doubtless help to account for the smaller number of female-headed businesses in the period. Indeed it is likely that many of these were run by widows who were carrying on their husbands' trade post-mortem (which was legally and socially more acceptable than a woman embarking upon and leading a business solely in their own name). The smaller scale nature of female-headed slopsellers also however points to particular difficulties of this trade. At a time when military demand for clothing was booming, it seems likely that women were held back by differential access to the credit necessary in this trade and by the nature of bidding for contracts which was often promoted by male forms of sociability.

In looking at Lemire's table (Table 4.9), it is clear that in the case of both male and female sellers, the mean is not a good indicator of average insured value. For women the mode would be the most appropriate value and, because the distribution is more evenly spread in the case of men, the median might be preferred. But remember, the choice of average is dependent upon the purpose of any research investigation and the median and mean are the only ones for which there is an accompanying set of measures of dispersion. One is prompted to ask of the table why there is no continuity in the range of insured values covered. Did this arise from some sort of sampling or from insurers only valuing to the nearest £100? It seems more likely that it is an error and that the categories should read £100–£299; £300–£499, and so on. One should also beware of the highest insured category which covers almost as great a range as the previous three categories combined. One would need to consult the original research to check on the range of values for men and women in this £1100 to £2000

History by Numbers

Table 4.9 Insured property of female and male slopsellers in London, 1777–1796

	Slop women	% of total	Slop men	% of total
£100–£200	11	29	31	17
£300–£400	14	37	60	34
£500–£600	6	16	34	19
£700–£800	2	5	13	7
£900–£1000	3	8	24	14
£1100–£2000	2	5	16	9
Totals	38	100	178	100

Source: Ms 7253 Royal Exchange Insurance Registers; Ms 11936&11937, Sun Fire Insurance Registers, Guildhall Library, London. Based upon table in Beverly Lemire, *Dress Culture and Commerce. The English Clothing Trade before the Factory, 1660–1800* (Basingstoke 1997), p. 51.

category. In neither case here, with such a small dataset and with uneven categories, would it be appropriate to undertake further statistical analysis of the distribution beyond the average, the frequency tables and the associated charts. As the figures are relatively easy to input into a spreadsheet, you may wish to draw up the relevant bar charts and derive frequency polygons for male and female insured values that will immediately highlight the (positively) skewed nature of each distribution (on the same graph).

Example 2: The impact of the Black Death in Birdbrook, Essex

In considering the local impact of the Black Death in Essex and specifically the impact upon tenurial developments and the availability of customary land, Phillip Schofield employed mean, median and standard deviation measures to demonstrate change over time. Table 4.10 shows mean length of leasehold where this indicates the period during which the tenement can be observed as remaining in the hands of the lessee by tracing it in the accounts from one year to the next. But the median is used to indicate the average term given at the inception of the lease to allow inclusion of terms granted for life or lives and to avoid replacing these with an arbitrary number of years. The standard deviation refers to leasehold lengths and relates to dispersal around the mean. The table shows that the average length of time that a lessee remained in his leasehold reduced dramatically in the first decade of the fifteenth century. The standard deviation of tenurial tenacity also declined markedly from the third quarter of the fourteenth century onwards. Schofield argues that these shifts reflected the replacement of a manorial economy based upon labour services with one based upon the money rent of farms and that a lot of the new tenants were incoming migrants.

Table 4.10 Average length of occupation leaseholds commencing in each decade, from 1350 to 1409

Decade	Mean length of leasehold ^a	Standard deviation	Median length of term ^b	Number of leases entered ^c
1350–9	17.5	13.162	12	6
1360–9	18	12.675	9	3
1370–9	11.3	11.609	7	7
1380–9	15.2	9.441	9	12
1390–9	10.6	5.795	3	13
1400–9	4	3.210	1	27

^a This is not the term given at the inception of the lease (see note b) but is the period during which the tenement can be observed as remaining in the hand of the lessee by tracing it through the accounts from one year to the next. Note also that the length of lease has been calculated as starting and ending in the first year of each account.

^b This is the term actually given at the inception of the lease, which, in the case of longer terms, would be recorded in the court roll or, in the case of very short terms, in the ‘farms’ section of the account. The median value has been used here rather than the mean so as to allow inclusion of terms granted for life or for lives without replacing these with an arbitrary number of years.

^c Three leaseholds entered in the decade 1380–9, four in that of 1390–9 and eight in that of 1400–9 had not expired by the accounting year 1409–10. Only limited observation is possible after this date: the next surviving accounts date from accounting years 1412–3, in which year the same lessees continue to hold, and 1426–7, by which date all but one of these lessees of customary tenements had disappeared. The mean length of leasehold has been distorted as a result: in the case of 13 of the 14 lessees still holding in 1409–10 it has been assumed, for the basis of the calculation, that their tenure of the lease ended in 1412–13, and the lease of the individual still *in situ* in 1426–7 has been taken as ending in that accounting year. The effect of this is, obviously, to reduce the size of the mean, but the accuracy of the trend can be tested by artificially extending the length of those leases whose terminal date cannot be observed. By adding 3 years after 1412–13 for those leases commencing in 1380–9, 7.5 years for those commencing in 1390–9 and 10 years for those commencing in 1400–9 the following means ad standard deviations are obtained:

Decade	Mean	Standard deviation
1380–9	15.75	10.248
1390–9	13.30	7.289
1400–9	7.26	7.214

Source: Phillip R. Schofield, ‘Tenurial developments and the availability of customary land in a later medieval community’, *Economic History Review*, 49, 2 (1996), p. 259.

Example 3: The impact of taxes and benefits on UK incomes in the late 1980s

In an example from more recent history, Figure 4.6 shows the effects of taxes and benefits upon quintile groups of households in Britain in 1987. It suggests that all five groups make direct and indirect tax contributions to the Welfare State and enjoy benefits in cash and kind. These taxes and benefits taken together make the distribution of final income considerably more equal than the distribution of original income.⁸

History by Numbers

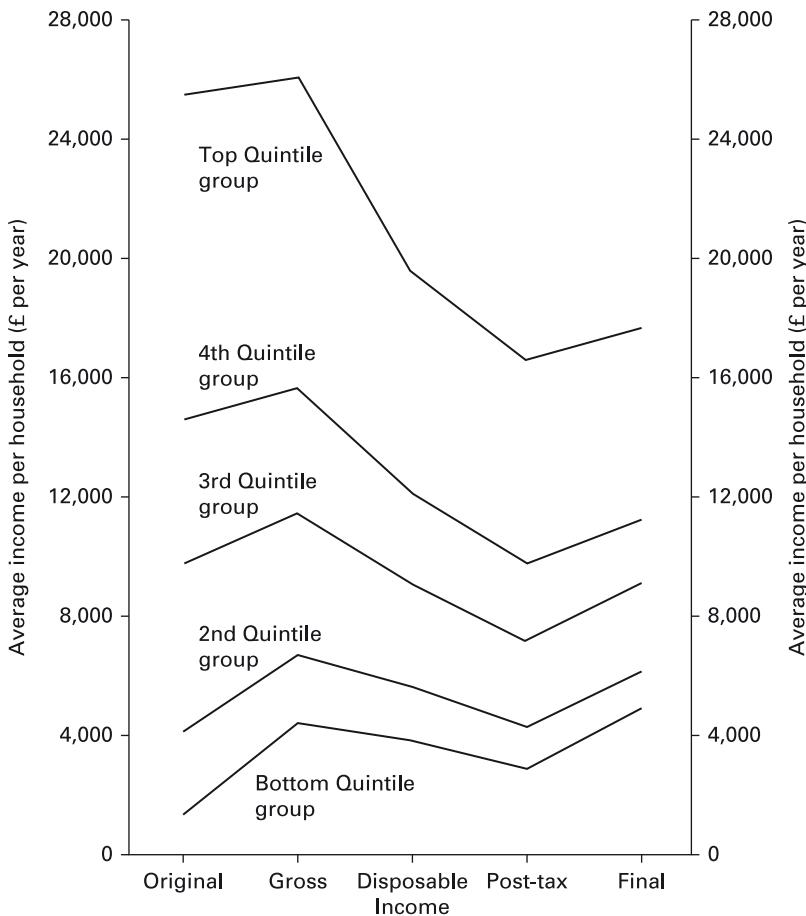


Figure 4.6 The effects of taxes and benefits on quintile groups of households, 1987.

Note: original income = employment and investment income before government intervention; gross income = original income plus cash benefits; disposable income = gross income minus direct taxes; post-tax income = disposable income minus indirect taxes; final income = post-tax income plus benefits in kind (e.g. health, education).

Source: Paul Johnson, 'The welfare state', in R. Floud and D. N. McCloskey (eds), *The Economic History of Britain Since 1700*, Volume 3, 1939–1992 (2nd edn, Cambridge 1994), p. 306. Based on *Economic Trends* (1990), no. 439, p. 88.

Quintiles are useful in Figure 4.6 in giving a clear idea of the differential effects of incomes, taxes and benefits across the spectrum of income distribution without clouding the diagram with an excessive amount of data that would add little to the point being made.⁹ What is actually being measured here are the **Gini coefficients** at each stage of the income/tax/benefits process. As explained earlier, the Gini coefficient is a summary measure of distributional equality between social groups. A Gini coefficient of 0 would denote absolute equality (the top 1 per cent and the bottom 1 per cent and all percentiles

Table 4.11 Gini coefficients for the distribution of income at each stage of the tax–benefit system, 1975–1987

Gini coefficients (%)	Year			
	1975	1979	1983	1987
Income type:				
original	43	45	49	52
gross	35	35	36	40
disposable	32	33	33	36
post-tax	33	35	36	40
final	31	32	33	36

Note: For definitions of income types, see Figure 4.6

Source: Paul Johnson, 'The welfare state', in R. Floud and D. N. McCloskey (eds), *The Economic History of Britain Since 1700*, Volume 3, 1939–1992 (2nd edn, Cambridge 1994), p. 305. Based on *Economic Trends* (1990), no. 439, p. 118.

in between each receive one per cent of total income). A coefficient of 100 indicates total inequality (the top 1 per cent receive all the income, the rest get nothing).¹⁰ The Gini coefficients relating to the data in Figure 4.6 are given in Table 4.11. They show that inequality grew between 1975 and 1987 and that this was a result of changes in original income (in turn affected by rising unemployment), rather than in the structure of taxes and benefits. Inequality has continued to grow on trend since the 1980s with the tax and benefit system having a much less positive impact on the distribution.¹¹ It is important to note that because Gini coefficients are independent of the original units of measurement (£, \$ etc.), they are very useful for cross-national and well as cross-class comparisons.

Example 4: The age of leaving home in the USA in the twentieth century

Good use is made of medians and ranges in a study of factors determining the age of leaving home in the US in the twentieth century, by Myron Gutmann and his co-authors,¹² in an article used as an exercise for readers later in this volume. Figure 4.7 is drawn from the article. It shows changes in the median age of leaving home by gender and race, between 1880 and 1990. The data is drawn from decennial census information about the numbers of young people remaining in the parental home so the information is derived from static benchmark evidence that has limitations. The statistical evidence is given in Table 4.12. Medians are here used instead of the mean or the mode because this is likely to best reflect average experience. The age range is generally fairly narrow but there is no clear mode over time and there are outliers (those who left home very early and those who stayed permanently) which would impact upon the mean measure.

History by Numbers

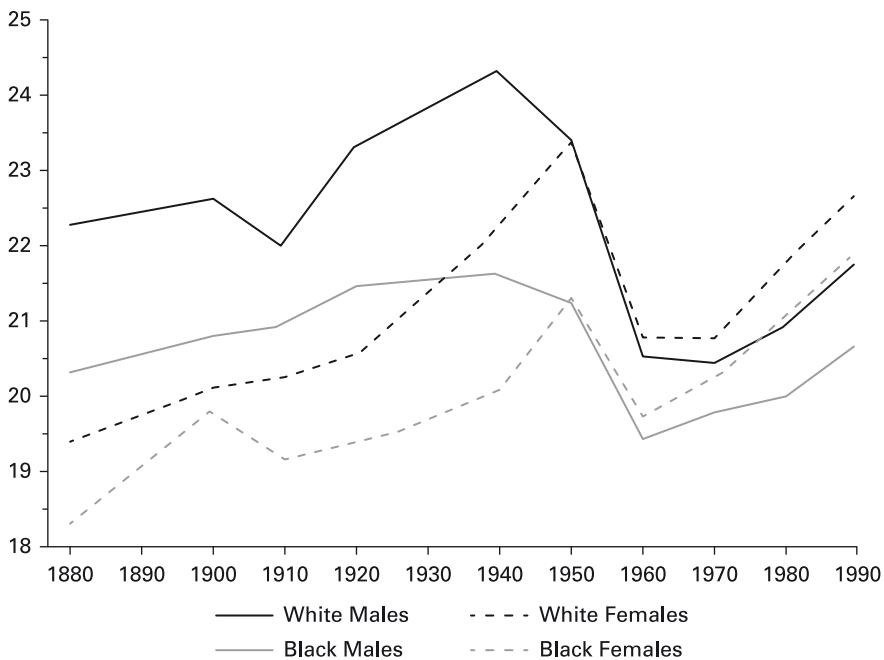


Figure 4.7 Median age at leaving home, United States, 1880–1990.

Source: M. Gutmann, S. Pullum-Piñón and T. Pullum, 'Three eras of young adult home leaving in twentieth-century America', *Journal of Social History*, 35 (3), (2002), p. 534. © Oxford University Press.

Figure 4.7 is the clearest way of demonstrating chronological change but it omits any indication of change in the range of these distributions, which is indicated in the estimations in Table 4.12. It is important to note that this table details the number of years lived with one or both parents rather than the age of leaving home because of the nature of the census evidence. Clearly in some periods and for some groups the experience of leaving home was more age-clustered than at other times. Can you spot these variations?

Studies of the age of leaving home, and the reasons for it, ideally require longitudinal evidence (such as detailed life histories) but this is rarely available for a sizeable or representative sample of the whole population so Gutmann et al. have calculated various probabilities that young people left home at particular ages based upon the ages at which most people were found co-residing with parents at the decennial census. It is not ideal but these simple statistical techniques have made it possible to use cross-sectional data to good effect. This sort of technique is called **Logit analysis**. Logit analysis was originally developed by the marketing industry to assess the scope of customer acceptance of a product, particularly a new product. It attempts to determine customers' purchase intentions and translates that into a measure of actual buying behaviour. In this case the major determinants of leaving home are compared

Summarizing Data: Averages and Distributions

Table 4.12 Estimated quartiles for number of years lived with one or both parents, United States, 1880–1990

	1880	1900	1910	1920	1940	1950	1960	1970	1980	1990
<i>White Males</i>										
1st Quartile	18.5	18.8	18.9	20.0	20.9	20.6	18.5	18.6	18.8	19.1
2nd Quartile (Median)	22.3	22.6	22.0	23.3	24.3	23.4	20.5	20.4	20.9	21.7
3rd Quartile	26.2	27.0	26.5	28.0	28.9	27.5	23.3	22.9	23.6	25.3
<i>White Females</i>										
1st Quartile	15.2	16.6	17.4	17.9	18.4	19.5	17.8	18.0	18.7	18.5
2nd Quartile (Median)	19.4	20.1	20.2	20.5	22.2	23.3	20.7	20.7	21.7	22.6
3rd Quartile	22.9	23.7	23.7	24.1	26.5	27.9	24.6	24.4	25.9	*
<i>Black Females</i>										
1st Quartile	17.2	17.6	17.9	18.5	18.8	18.8	17.8	18.1	18.2	18.5
2nd Quartile (Median)	20.3	20.8	20.9	21.4	21.6	21.2	19.4	19.7	19.9	20.6
3rd Quartile	24.6	25.2	25.4	26.1	26.2	24.8	21.5	21.8	22.1	23.4
<i>Black Females</i>										
1st Quartile	15.9	16.2	16.6	16.9	17.1	18.0	17.2	17.8	18.4	18.5
2nd Quartile (Median)	18.3	19.8	19.1	19.3	20.0	21.3	19.7	20.2	21.0	21.8
3rd Quartile	21.3	22.8	22.6	22.6	24.3	27.1	23.1	23.3	24.6	27.2

*estimate unavailable

Source: M. Gutmann, S. Pullum-Piñón and T. Pullum, 'Three eras of young adult home leaving in twentieth-century America', *Journal of Social History*, 35 (3), (2002), pp. 533–576, p. 558. © Oxford University Press.

chronologically with the number, age and nature of those co-resident in the census data and this is used to infer real decisions about leaving home and the changes in the age of leaving home for different gender and racial groups. You can test your understanding of this article and the statistical techniques employed, as well as learning about the manifold and changing determinants of leaving home (education, orphanhood, military service, employment prospects, wage levels, changing age and importance of marriage and first cohabitation, immigration status and so on) by doing the exercise in the section following Chapter 7.

DISTRIBUTIONS

We have seen that distributions can cover a very wide range of values or they can be made up of numbers that are clustered closely together. Distributions also take on different shapes, tending towards symmetry, or a skew.

The normal distribution

There is an ideal-type of distribution, known as the **normal distribution**, which is used in statistical theorizing. The expression ideal-type is generally used to indicate a phenomenon which does not occur exactly in practice but has characteristics commonly found in real phenomena.¹³ Thus normal distributions rarely occur exactly in social or historical data but in large-scale distributions and especially in the natural sciences the binomial, bell-shaped distribution is the one to which real data distributions often tend (as in Figure 4.2 which gave the heights of US passport applicants). In the normal distribution the mean, the median and the mode have the same value with an equal number of observations spread out symmetrically on either side. The normal distribution, as we shall see in Chapter 7, is also the basis of sampling theory in statistics. It is thus useful to know about the properties of the normal distribution.

In a normal distribution a constant proportion of cases lie between the mean and multipliers of the standard deviation from the mean:

68.26 per cent fall between one standard deviation above and below;

95.46 per cent fall between 2 standard deviations above and below;

99.7 per cent fall between 3 standard deviations above and below.

The **normal distribution** can be represented graphically as shown in Figure 4.8.

The bell-shaped curve of the normal distribution underlies much theorizing about statistics and probability. The ‘average man’ whose characteristics were described by Adolphe Quetelet (1796–1874) was conceived and recorded in this way in all his physical attributes:

After his study of heights, Quetelet continued his measurements of other physical attributes: arms and legs, skulls and weights, for which he still observed distributions in accordance with binomial law. From this he inferred the existence of an ideal average man, in whom all average characteristics were combined and who constituted the Creator’s goal – perfection.¹⁴

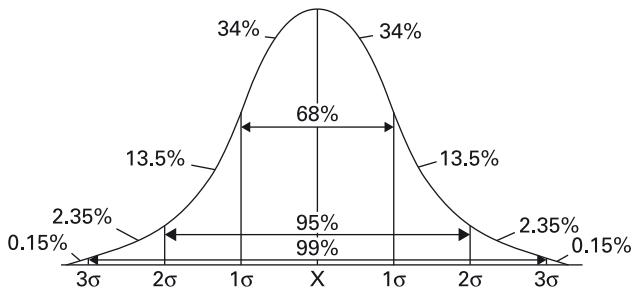


Figure 4.8 The normal distribution.

Note: σ = standard deviation; X = mean and mode.

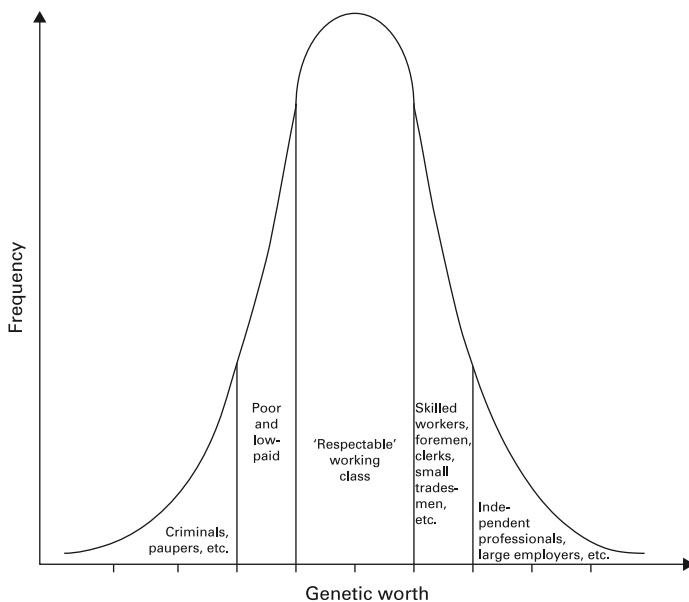


Figure 4.9 Social classes and genetic worth (Galton 1909).

Source: Alain Desrosières, *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, MA 1998), p. 114.

The ‘perfection’ approached by Quetelet in this way was the normal or binomial distribution. Galton was influenced by Quetelet, by the social investigations of Charles Booth and by Darwinian theories of evolution in theorizing the distribution of ‘genetic worth’ as a normal curve (see Figure 4.9).

Skewed distributions

Other distributions commonly occur with historical and social data where the spread of observations is uneven with more lying either above or below the mean. Where most observations lie below the mean the distribution is described as **positively skewed** (Figure 4.10a). Where most observations lie above the mean the distribution is described as **negatively skewed** (Figure 4.10b). These distributions are represented graphically below with the relative positions of the mode and the median as well as the mean indicated. It is easy to see why the mean is not always a good measure to use for the average of a **skewed distribution** and it is usual in these cases to give the value of all three averages.

The distribution of the land tax payers of Sowerby given in Table 3.18 is skewed in favour of those paying under £5 and could be roughly drawn as shown in Figure 4.11.

Another example of a skewed distribution is the age profile of cotton workers in the nineteenth century as indicated in Figure 4.12. Young people were favoured as employees with the female workforce appearing particularly youthful. There was however a long ‘tail’ of older workers.

History by Numbers

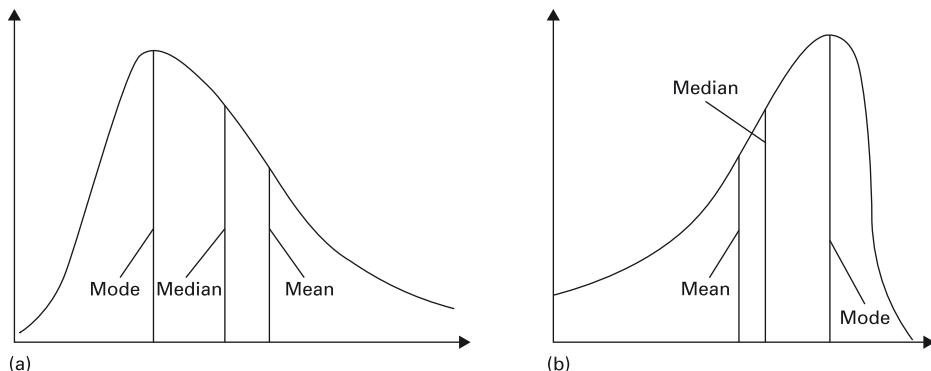


Figure 4.10 Skewed distributions: (a) positive skew (mode and median less than mean); (b) negative skew (mode and median more than mean).

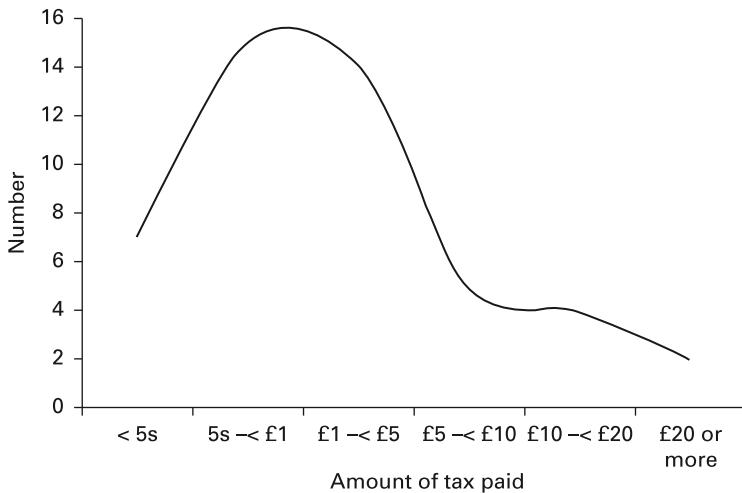
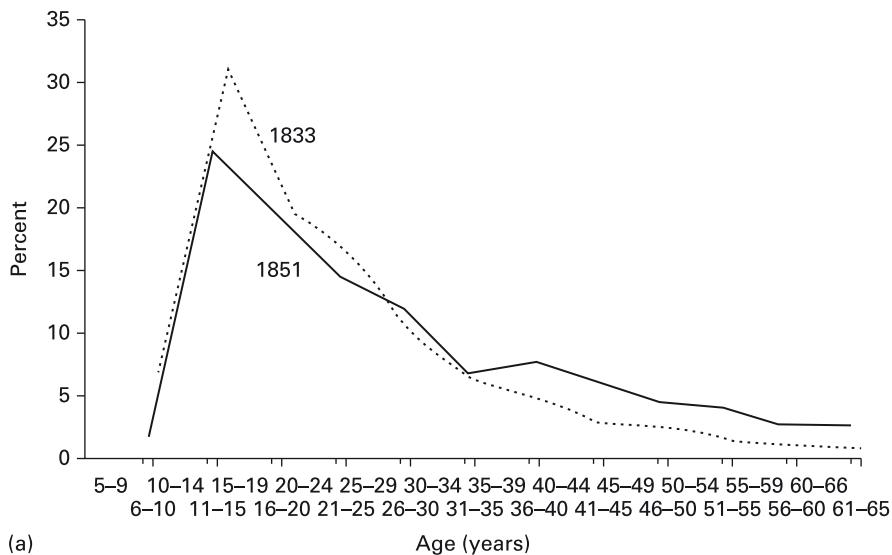


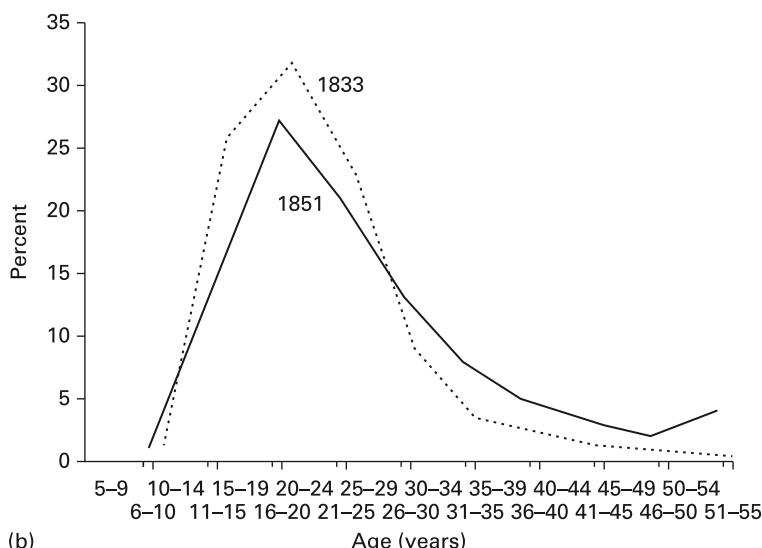
Figure 4.11 Distribution of land tax payers, Sowerby, West Yorkshire, 1782.

Source: see Table 3.18.

In our example in Chapter 3 of prisoners in Portland Prison in 1849 sentence lengths have a negative rather than a positive skew. The mean (13.5) is lower than the median (14) or the mode (15) (Figure 4.13).



(a)



(b)

Figure 4.12 Age distribution of the Lancashire cotton industry workforce in 1833 and 1851: (a) males; (b) females.

Source: H. M. Boot, 'How skilled were Lancashire cotton factory workers in 1833?', *Economic History Review*, 48, 2 (1995), p. 286.

History by Numbers

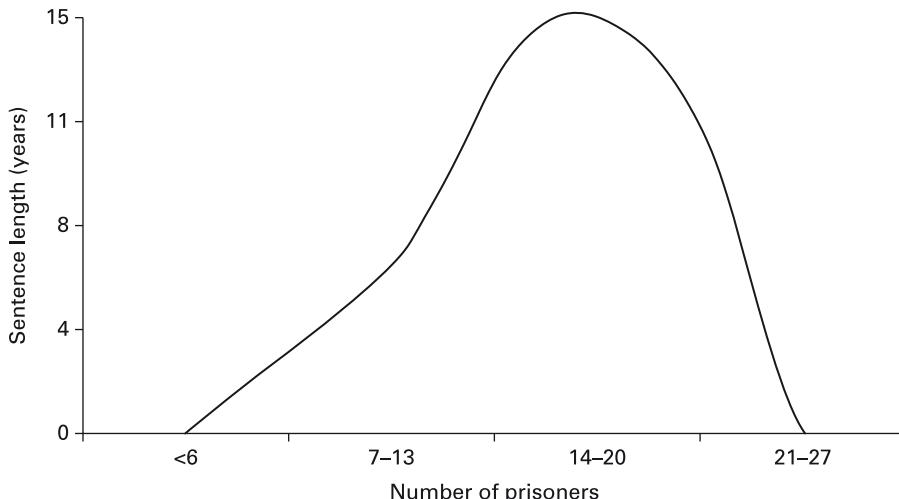


Figure 4.13 Distribution of sentence lengths, Portland Prison, 1849.

Source: Table 3.12.

Distributions with more than one mode

Sometimes distributions occur where there is more than one value around which observations cluster. The occurrence of such distributions illustrates the importance of studying the distribution carefully and perhaps graphing it or drawing a histogram or frequency polygon before rushing to select and calculate an average measure. Figures 4.14 and 4.15 show bi-modal and tri-modal distributions respectively.

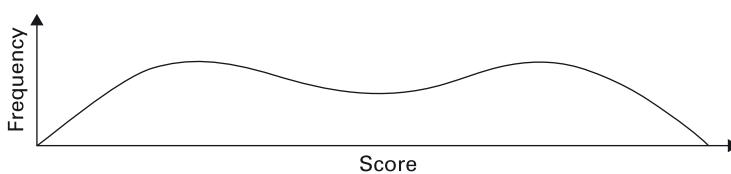


Figure 4.14 Bi-modal distribution.

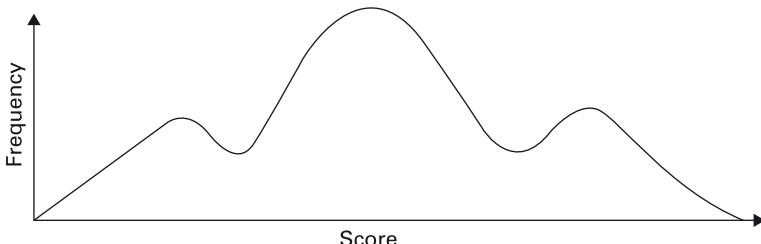


Figure 4.15 Tri-modal distribution.

Conclusion

The most common piece of elementary statistical analysis involves summarizing and considering the nature of a distribution or distributions of values. Measures of central tendency and of dispersion, together with the possibilities presented by graphing the distribution go a long way toward making sense of data and enabling one to compare one distribution with another. These calculations and techniques are important in themselves but also as a preliminary to further, more sophisticated, analysis.

Further reading

- Daly, F., D. J. Hand, M. C. Jones, A. D. Lunn and K. J. McConway, *Elements of Statistics* (Harlow 1995), Chapter 1.
- Feinstein, Charles, *Making History Count: A Primer in Quantitative Methods for Historians* (Cambridge 2009).
- Foster, Liam, Ian Diamond and Julie Jeffries, *Beginning Statistics: An Introduction for Social Scientists*, 2nd edition (London 2015), Chapters 4–7.
- Gonick, Larry and Woollcott Smith. *The Cartoon Guide to Statistics* (New York 1993).
- Hanagan, T., *Mastering Statistics*, 3rd edition (London 1997), Chapters 4 and 5.
- Haskins, Loren and Kirk Jeffreys, *Understanding Quantitative History* (Cambridge, MA 1991), Chapters 1–2.
- Solomon, R. and P. Winch, *Calculating and Computing for Social Science and Arts Students* (Buckingham 1994), Chapter 4.
- Tufte, E., *Visual Explanations: Images and Quantities, Evidence and Narrative* (Cheshire, CT 1997).

Exercises for Chapters 3 and 4

Moreton, Emma, 'Profiling the female emigrant: a method of linguistic inquiry for examining correspondence collections', *Gender & History*, 24 (3), (2012), pp. 617–646.

1. Why might historians think that emigrant letters are a useful source for historical gender analysis?
2. What are the benefits of a quantitative approach over qualitative analysis of emigrant letters and what are the shortcomings?
3. What is a corpus and how does it differ from a digital archive?
4. Define 'corpus linguistics' as used here.
5. Why is comparative data central to the technique of quantitative analysis of corpus data?
6. Comment on the use of Antconc in relation to similar softwares.
7. Why is the term token used?
8. What advantages does the arrangement of information into Tables 1–3 have over the explanation given in prose? Are there any drawbacks or pitfalls of this rearrangement of data?
9. What is the purpose of calculating a type/token ratio in this article? How safely can one draw conclusions from this?
10. Why does the author 'normalize' the raw data of the frequency of common words, in Table 10 for example?
11. What is a concordance line and why is it used?
12. Why is the frequency of particular verbs so important in this analysis?
13. Why is clustering around the verb a focus of attention?
14. How easy is it to make a shift from quantitative analysis to qualitative insights in this article?
15. How convincing is the evidence of certain repeated phrase patterns for the hypothesis that a) familial relations of emigrants were strengthened and reinforced by letter writing and b) that men and women wrote and used letters and words in different ways?
16. Can the method used here be applied to other historical sources? If no, give reasons. If yes, suggest some examples of research you would be interested in reading or doing using the method.

Desrochers, Robert E., Jr, 'Slave-for-sale advertisements and slavery in Massachusetts, 1704–1781', *William and Mary Quarterly*, 59 (3), (2002), pp. 623–664.

1. Comment on the fact that the study's chronology is determined by the availability of the advertisement sources?
2. Comment on the role of Boston in the slave trade.
3. Why does the author give such a detailed description of Boston's geography on pp. 627–629?
4. Speculate on the gender difference between male and female slave advertisements.
5. Regarding Tables I and II, what are the problems of the presentation of the tables? How can they be improved for greater clarification?
6. Regarding the increasing anonymity of slave sellers, can you give an alternative explanation to the author's interpretation? Can it be interpreted as the institutionalization of slave sales?
7. Why did the adverts mention the origin of slaves and the years spent in the Americas?
8. How did the author compute Table V? What problems are there in this table?
9. Compare the slaves-for-sale adverts used in this study with the runaway slaves adverts and consider the accuracy of the information about the slaves included in the two types of advertisements. South Carolina runaway slaves sources: <http://libcdm1.uncg.edu/cdm/landingpage/collection/RAS>.
10. Does the fact that a slave could be owned by more than one owner pose a problem when using the adverts?
11. What do the slave adverts tell us about the desirability of slaves in Boston in terms of language, origin and health?
12. Choose two years from Table VI and draw pie charts to best convey the information.
13. What was the scale of slave imports in the Massachusetts area compared to South Carolina?
14. How did the author compute the number in the 'slave per advertisement' column in Table VIII? Could Tables VII and VIII be merged? Would this have been more effective?

Dunstall, Graeme, 'Frontier and/or cultural fragment? Interpretations of violence in colonial New Zealand', *Social History*, 29 (1), (2004), pp. 59–83.

1. What is the purpose of this piece of research?

History by Numbers

2. Why might historians be keen to question the long-term U-shaped trend for violent crime in Western societies?
3. Why are moving averages used in Figures 1 and 4?
4. Why are crime charges per 100,000 used on the vertical axes of many of the figures?
5. Why might the homicide rate in the late nineteenth and early twentieth century be considerably altered by excluding female perpetrators and 'accidental' or involuntary cases?
6. What aspects of 'frontier' and 'settler' society might have predisposed such societies to violent crime?
7. What social and cultural changes of the late nineteenth century in New Zealand might account for declining violence?
8. Why might alcohol and its control be such an important determinant of shifts in violence and to what extent is this supported by the statistical evidence?
9. What does the evidence of gender-specific violent crime, exhibited in Figures 7 and 8 suggest about the 'atomisation thesis'?
10. What may have been the role of colonial litigiousness (and its determinants) in explaining the long-term trends in violent crime?
11. Why were so many prosecutions for assault withdrawn in the late nineteenth century?
12. How might the data in Figure 9 be interpreted? Would it help to know what sorts of punishments resulted from convictions and if they changed over time?
13. To what extent does the key argument about the impact of policing in the late nineteenth century in England and Wales hold for New Zealand?
14. Why might further research in Middlesborough or Merthyr Tydfil support the New Zealand-specific explanations of crime trends?

French, Michael, 'Commercials, careers, and culture: travelling salesmen in Britain, 1890s–1930s', *Economic History Review*, 58 (2), (2005), pp. 352–377.

1. In the case of commercial travellers why is it difficult to interpret evidence about their income?
2. What is the advantage of using indices as measures? (pp. 356ff)
3. In Figure 1 (p. 357), how has the average real income index for McVitie travellers been established and what are the main pitfalls of this exercise?
4. In Figure 2 the average real incomes of Coats travellers are given as 1889–1929. How easy is it to compare Figures 1 and 2 for the time period when they overlap? What would make it easier?

5. Using the data from Figure 3 take two contrasting years and draw up income frequency distribution tables. Comment on the causes of the different distributions.
6. From Table 1 take four workers and use a different method of displaying the movement of their relative salaries, 1922–1937.
7. Many figures are dotted around in the text of this article. Are there any other points at which you feel that the author might have used a better and more visual method of conveying the data?
8. How well do you think that the author integrates the quantitative and the qualitative information in the article? Where might it have been improved?

Vickers, Daniel and Vince Walsh, 'Young men and the sea: the sociology of seafaring in eighteenth-century Salem Massachusetts,' *Social History*, 24 (1), (1999), pp. 17–38.

This article uses only simple descriptive statistics and is a rare application of quantitative methods to research using life histories. It is a good exercise for those just starting to feel their way with quantitative work.

1. What sorts of evidence and research techniques have been used in this piece of research?
2. Does nominal record linkage appear to have been undertaken?
3. Why did young Salem men go to sea?
4. How are the proportions of Salem men who were seafarers calculated? (p. 24)
5. Would the random sample mentioned in footnote 22 guarantee an unbiased sample of Salem Ship Masters? (For a definition and explanation of random sample see p. 208 of this volume.)
6. Comment on the use of deciles on p. 25.
7. Is there any better way of representing all of the information contained in Figure 1?
8. Why has the median age of first recorded voyage been used in footnote 29 rather than the mean or the mode?
9. What advantage has the pie chart on page 27 got over alternative ways of presenting the data on the fate of young seafarers?
10. What variables appear to have been important in determining which seafarers were promoted to mate or master?
11. What is meant by 'age-specific mortality' and the 'population at risk' in footnote 30?
12. Why has the mean age of marriage been used in footnote 42?

13. For a small community, Salem appears to have received the attention of several researchers in the early twentieth century which these authors have been able to benefit from. What problems might attach to heavy reliance upon the works of Tapley, Perley, Crandall and Essex Institute?
14. Why have the contrasting median ages of Salem-born and other seamen been relegated to footnote 62 and why has the median been used here?

Mitchell, David M., 'My purple will be too sad for that melancholy room: furnishings for interiors in London and Paris, 1600–1735', *Textile History*, 40 (1), (2009), pp. 3–28.

1. To what extent is the time period addressed here determined by the cover of the evidential base? Are there other justifications? Why are these important?
2. Discuss the advantages and the limitations of using probate inventories for this sort of investigation.
3. Why compare findings from interiors in the Paris study?
4. What difficulties are involved in using data generated by another project as a comparator?
5. What distortions may have been introduced into the evidence by the choice of 15-year periods and the particular room categories selected?
6. Which of the figures in the article is a frequency polygon?
7. How many figures are bar charts and how many are in the form of histograms?
8. What results in Tables 3 and 4 suggest the introduction of bias via the number and sort of inventories included over time?
9. Comment on the possible impact of variation in sample size upon the results displayed in Figure 5.
10. Choose a room value range and draw a bar chart showing the distribution of fabrics in bed hangings and other room hangings, based on the data in Tables 1 and 5. Make sure you adopt a professional style of heading and an appropriate source note.
11. Look over the various tables and figures and comment on their professional presentation. Is there anything missing from Figure 7 for example?
12. With reference to Figures 7, 8 and 9, comment upon the social and chronological variation in colour preferences expressed.
13. Given the comparative data for Paris, as well as that for London, how successful is Mitchell's claim that taste played more of a role in colour choice for interiors than supply side factors such as improvements in dyeing?

14. How well does Mitchell marry the quantitative and the qualitative evidence? What difference does the qualitative evidence in pp. 21–24 make to the conclusions?

Owens, Alastair, 'Property, gender and the life course: inheritance and family welfare provision in early nineteenth-century England', *Social History*, 26 (3), (2001), pp. 299–317.

1. This study is based upon analysis of a sample of wills for Stockport, 1800–1857. Why might the information in footnote 11, and in the text on p. 302, be seen to be inadequately reassuring about a) the way in which the sample of wills was selected and b) about the extent to which the wills studied represent property and possessions passed on to children in Stockport in this period?
2. The study concentrates upon gender difference in 'the transmission of urban industrial fortunes' (p. 303). Suggest one major shortcoming in addressing this issue that is not directly covered by the author. Explain your choice.
3. What sort of statistical software and technique of analysis might have assisted in the analysis and demonstration of gendered discourse in will writing, had all of the wills been available in machine-readable format.
4. Using the information in the bar chart in Figure 1 (transferring it to spreadsheet software if you wish), construct another visual method of clearly displaying the same information.
5. Critically discuss the weaknesses of the information provided in Table 1, including the 'Total' vector.
6. Using matrix notation indicate from Table 1 the location of the number of real estate bequests to daughters and the number of investment bequests to sisters.
7. With regard to Figures 2 and 3 comment on the use of the term probability rather than percentage in the titles of these figures. Which would be best and why?
8. Using the information from Table 2 (transferring it to spreadsheet software if you wish) create a more visual way of displaying this comparative data.
9. Discuss the advantages of having used some descriptive statistics in this study. Might more use have been made?

Shepard, Alexandra, 'Crediting women in the early modern English economy', *History Workshop Journal*, 79, (2015), pp. 1–24.

1. This journal does not include abstracts at the start of each essay. Write a 200 word (max) abstract for this article, explaining what it seeks to do and why but also highlighting the rearrangement and display of the quantitative evidence.

History by Numbers

2. Why are witness depositions a good source for considering the economic position of women in early modern England?
3. Looking at Table 1, after reading the whole article, should we be wary of how the geographical range and the variability of surviving depositions might influence overall results?
4. What other biases should the historian look out for in using witness depositions as a source for the sorts of work people did?
5. How reliable is the information in Table 4 as a guide to change over time in living by one's labour for men and for women?
6. Using the material in Table 5 draw a bar chart to represent the same information.
7. Using the information in Table 6 draw three pie charts to represent the same.
8. How wary should we be about accepting the high level of concentration of single women in domestic service? What factors are likely to have contributed to this high concentration?
9. Using the categories of 'making and mending clothes' and of 'textile manufacture', give probable reasons for the differences between the different participation of women in London and outside in these textile employments.

Young, David M., 'Social democratic federation membership in London', *Historical Research*, 78 (201), (2005), pp. 354–376.

1. The database of SDF members built by the author contains over 1,500 names. From what the author describes as his sources for this (p. 355) and from what one learns later in the article, comment on the most likely types of bias that might arise purely from the selection of members included in the database.
2. Why are the two vertical scales useful in the Figure 1 graph?
3. Why did so many SDF branches have a short life?
4. Why was there such a high turnover in branch membership?
5. Use the evidence in Figure 2 to create:
 - a frequency table of the membership length of the 122 members;
 - a percentage frequency table of membership length;
 - a cumulative percentage frequency table of membership length.

Which way of displaying this data do you think is the most effective?

6. Using the information in Appendix B, devise two other alternative effective ways of displaying the occupational data to that of the flat pie chart in Figure 3.
7. Display in a new figure and in greater detail than the author manages, the occupational composition of the skilled/craft SDP members in his database.

8. Which of Figures 4a and 4b do you feel is most effective in showing the age of political activity and why?
9. Does Figure 6 (p. 368) provide a reassurance for the elements of Figure 4?
10. What is the most likely cause of the steep drop in years of activity after the first year or two indicated in Figure 6?
11. What other information would you most like to see alongside the average age graph in Figure 5? What are the main pitfalls of the information in Figure 5 as an indicator of the most common age of political affiliation or action? (List at least 3.)

Menard, Russell R., 'Making a "popular slave society" in colonial British America', *Journal of Interdisciplinary History*, 43 (3), (2013), pp. 377–395.

1. Comment upon the strengths and weaknesses of probate records as a source for contributing to the debate about 'slave society' in the Chesapeake.
2. The Tables in the article show different ways of arranging the evidence in the probate inventories in order to address the questions posed. Discuss the strengths and weaknesses of these rearrangements and the ways in which the data in the Tables are headed and displayed. (Refer particularly to Chapter 3 for guidance here.)
3. In Tables 1 and 2 which wealth category showed most drastic change in slave owing between the early- and mid-eighteenth century?
4. In what ways are the conclusions drawn from Table 3 about life cycle ownership of slaves open to question?
5. How meaningful are the figures in Table 4 for households with more than 20 slaves?
6. What useful additional column might have been added to Table 5 to show shifts in major slave-owning neighbourhoods between the early- and mid-eighteenth century?

Dobson, Stephen and John Goddard, 'Performance revenue and cross subsidisation in the Football League, 1927–1994', *Economic History Review*, 51 (4), (1998), pp. 763–785.

There is no complex quantification here but the article provides a good argument where the (somewhat clumsy) descriptive statistics are essential to the case.

1. With what justification can the Football League as a whole be regarded as a firm?
2. Why do sporting authorities commonly restrict player mobility, decide who plays whom, when and where, and promote the sharing of gate receipts, commercial and media revenues?

History by Numbers

3. In Table 1 what is the coefficient of variation and why is it given here?
4. What are real gate revenues and why are they given (alongside real average admission prices) in index form?
5. Could the data in Table 2 have been more attractively and convincingly displayed?
6. Are there any difficulties with the way in which the performance score has been derived and in comparing this score with attendance and gate revenues in Table 3?
7. Could the data in Table 4 have been more attractively and convincingly displayed?
8. What was the rationale and the effect of the maximum wage and the retain and transfer system that existed until the early 1960s?
9. Could the information in Table 5 have been more effectively displayed for the purposes of the argument being made?
10. What was the impact of the overhaul of the retain-and-transfer system in the 1960s and in 1977?
11. In what ways has the transfer system since the 1960s resulted in a redistribution of income between clubs?
12. Could the data in Table 6 be more attractively and convincingly displayed?
13. What has been the impact of the erosion of explicit schemes for the redistribution of match receipts between clubs and how did cup competitions help to compensate for this?
14. What was the financial effect on the rest of the Football League of the creation of the Premier League?
15. Why are the authors sanguine about the rise of increasingly individualistic, profit-motivated attitudes amongst the major clubs in particular?

Solar, Peter M., 'Opening to the East: shipping between Europe and Asia, 1770–1830', *Journal of Economic History*, 73 (3), (2013), pp. 625–661.

1. What is the purpose of this article?
2. What argument is the data in Figure 1 used to support and how?
3. Discuss the amalgamation of different sources used to create Figures 2 and 3 and explain why this has been necessary. How accurate are the composite figures?
4. Why did ships become smaller in the Asia trade over time and what effects did this have?
5. How convincing is the data on increased frequency of voyages of individual ships over time? How well is this data displayed for the purposes of the argument?

6. Why might ships have been used more intensively?
7. Discuss the relevance of the findings presented in Figure 5.
8. How well does the author evidence and explain the changes in shipping costs on the Asian routes?
9. Why is it important for the author to demonstrate that changes in shipping had a measurable impact upon the prices of imported goods and how convincing are his data on this?

CHAPTER 5

TIME SERIES AND INDICES

As Donald Coleman proclaimed, 'For the historian time is of the essence; it embraces growth and decay, stagnation or adaptation, change in all its complexity'.¹ Historians are most often concerned with change over time and with chronological variation. They thus commonly need to collect and to consider statistical data arranged in **time series**.

A **time series** consists of numerical data recorded at intervals of time in chronological order. It is thus a special case of a data vector in which measures of a variable are in chronological order. Time intervals can be daily, weekly, monthly, yearly or by groups of years, for example each quarter or half-century. As long as the intervals are regular, there are statistical techniques that can be used to analyse the series. The variable altering over time (for example, exports, wages, capital investment, strikes, crimes, births, deaths) can be expressed as a monetary value, volume or quantity or it might be converted into an **index** (also known as *ratio value*), as explained below.

Figure 5.1 showing the death rates of mineworkers in various countries in the later nineteenth century is an example of comparative data collected in time series. The

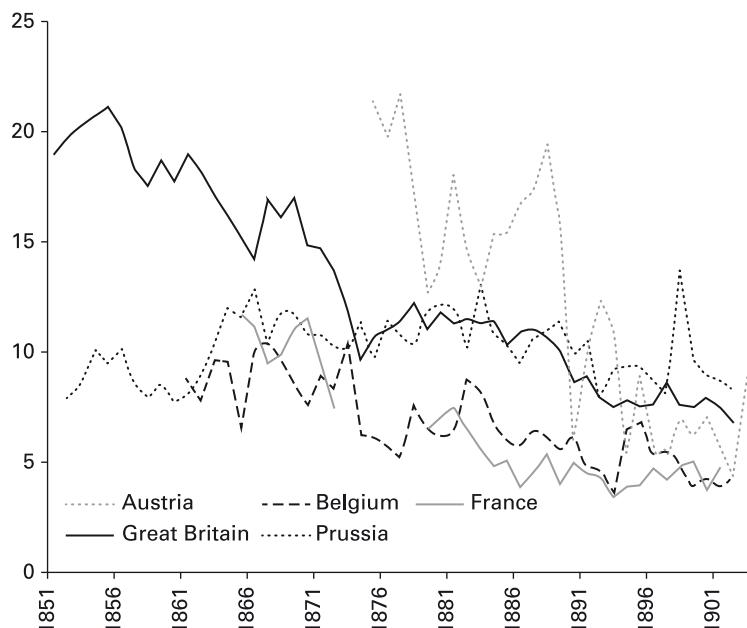


Figure 5.1 Death rate per 10,000 worker years due to mining accidents in five European countries, 1851–1901.

Source: John E. Murray and Javier Silvestre, 'Small scale technologies and European coal mine safety, 1850–1900', *Economic History Review*, 68 (3), (2015), p. 899.

History by Numbers

periodicity is not given but it looks like annual data (although the horizontal axis is in five-year intervals). It is immediately easy to compare change over time in the five countries as well as overall. Note that the death rate is per 10,000 worker years. This is a standardizing tool that allows for direct comparisons of relative safety taking into account the size of the mining workforce in different countries and different working hours.

Note that the original title for this graph in the article was ‘Death rate per 10,000 worker years due to roof collapses, falling earth, cave ins and so on’. This was not a good choice of title as it omits indication of geographical comparison and any sense of the period covered. Although this might be obvious from the graph itself, if one is writing a book or dissertation it is particularly important to convey as much essential information as possible in the title of a table or figure so that when the title is listed in the contents of the volume it is immediately easy to judge what it may contain.

The rest of this chapter considers the prospects and pitfalls of the collection, manipulation and analysis of time series.

Index numbers (indices)

An index number (or ratio value) is the value of a variable expressed as a percentage. The percentage is calculated as a proportion of the value which the variable holds in a so-called **base period** (most often a **base year**).

What are the advantages of using index numbers (or indices) rather than original data?

- Indices enable easier identification of trends and variations in the time series especially where the original units are complicated, for example, £. s. d., tons and cwts, bushels, acres, rods and perches.
- Indices make it easy to compare the movement of two or more simultaneous time series one with another, if they have the same base year. This is especially important where the original data values of the different series have different units, for example strikes and average wages, beer in millions of gallons produced and tobacco consumption in lbs per head.
- Index numbers enable the formation of **composite and real indices** (see below) which express the movement of weighted or adjusted variables.

The formation of indices

Indices are generally formed for prices, quantities or monetary values. To convert a time series from original values to index numbers it is first necessary to select a base year (month or day – depending on the time intervals involved). The index value of 100 is given to the data value for that year. Every other year is then expressed as a percentage of the base year.²

In Table 5.1 1866 was chosen as the base year. The index for 1865 ($I^{(1865)}$) is then calculated as follows.³

$$\begin{aligned} I^{(1865)} &= \frac{\text{number of strikes in 1865}}{\text{number of strikes in base year (1866)}} \times 100 \\ &= \frac{45}{30} \times 100 \\ &= 150 \end{aligned}$$

Similarly, the index for 1861 is calculated as:

$$\begin{aligned} I^{(1861)} &= \frac{23}{30} \times 100 \\ &= 77 \end{aligned}$$

and so on. Note that once a base year has been chosen it should be indicated somewhere in the heading of the table or vector. In their early computer-aided study of strikes in France, Edward Shorter and Charles Tilly used a mass of time series variables, often converted to indices, in this way, order to conclude that strikes had changed over the period to become larger, more frequent and shorter in duration.

Table 5.2 gives beer output in barrels and tobacco consumption per head in lbs between 1925 and 1938. These are difficult to compare at face value because they are expressed in very different units. They are much easier to compare if we convert both series into indices with the same base year. Once this is done the remarkable feature that shows up in the indices is the extent to which the consumption of tobacco per head of the population held up during the high unemployment years of the 1930s compared

Table 5.1 Strikes in France and index of strikes, 1860–1870

Year	No. of strikes	Strike index (1866=100)
1860	20	67
1861	23	77
1862	22	73
1863	26	87
1864	29	97
1865	45	150
1866	30	100
1867	33	110
1868	35	117
1869	36	120
1870	38	127

Source: Hypothetical data patterned on work of Edward Shorter and Charles Tilly, *Strikes in France 1830–1968* (Cambridge 1974).

History by Numbers

Table 5.2 Beer output (millions of barrels), tobacco consumption (pounds weight per head) and net income (pounds sterling per head), 1925–1938

Year	Beer output	Tobacco consumption	Income	Indices (base year 1929)		
				beer	tobacco	income
1925	26.8	2.96	88.2	107	91	96
1926	25.2	3.00	86.6	100	93	95
1927	25.4	3.04	91.3	101	94	100
1928	24.6	3.11	91.1	98	96	100
1929	25.1	3.24	91.4	100	100	100
1930	23.9	3.31	86.2	95	102	94
1931	20.8	3.27	79.5	83	101	87
1932	18.0	3.23	77.1	72	100	84
1933	20.2	3.22	80.2	80	99	88
1934	20.9	3.41	83.1	83	105	91
1935	22.0	3.51	87.6	88	108	96
1936	22.7	3.72	93.2	90	115	102
1937	24.2	3.87	97.6	96	119	107
1938	24.7	4.00	98.3	98	123	108

Source: B. R. Mitchell, *British Historical Statistics* (Cambridge 1988), pp. 709–711, 829.

with beer output which remained consistently lower than the base year output in 1929. Of course one immediately wonders whether the movement of beer output is likely to reflect beer consumption per head. If so an interesting comparison could be made about the impact of the conditions of the 1930s upon beer and tobacco consumption. Lower beer output figures might however reflect the loss of export markets and this would have to be checked from other evidence before one could be sure that the output reflected domestic consumption habits. The distribution of domestic consumption amongst different social classes as well as average consumption per head might also be important in assessing the findings.

In Table 5.3, drawn from a recent book *British Economic Growth 1270–1870*, national output, labour force and output per worker figures have all been converted to the same base year of 1700 to allow for immediate comparison although note should be taken of the change in observations from England to Great Britain that occurred in that year. This impacts upon any straightforward comparisons pre- and post-1700.⁴ A useful exercise to test your skill and understanding of base years after reading this chapter would be to convert the observation for some or all of the years and categories in this table to base year 1851. What is the effect on the appearance of the series of so doing (see below)?

Converting raw data into indices with the same base year also enables them to be graphed together for comparative purposes. Figure 5.2 is reproduced from Charles

Table 5.3 Indexed trends in output, labour force and output per worker, England, 1381–1700 and Great Britain, 1700–1851

Sector	1381	1522	1700	1759	1801	1851
A. Output						
Agriculture	50.9	51.3	100.0	159.2	227.0	328.3
Industry	18.9	27.6	100.0	144.7	275.2	1,206.3
Services	24.8	27.1	100.0	150.9	266.6	777.4
GDP	29.2	34.2	100.0	150.4	251.6	711.5
B. Labour-force						
Agriculture	68.7	64.7	100.0	114.2	137.1	188.2
Industry	26.3	31.3	100.0	120.3	180.3	428.0
Services	40.5	34.8	100.0	130.0	197.2	404.5
GDP	46.6	45.2	100.0	120.6	168.1	328.4
C. Output per worker						
Agriculture	74.2	79.2	100.0	139.5	165.6	174.4
Industry	71.8	88.4	100.0	120.3	152.7	281.9
Services	61.3	78.0	100.0	116.1	135.2	192.2
GDP	62.6	75.7	100.0	124.7	149.7	216.6

Source: Stephen Broadberry, Bruce M. S. Campbell, Alexander Klein, Mark Overton and Bas van Leeuwen, *British Economic Growth 1270–1870* (Cambridge 2015), p. 365.

Feinstein's article on the standard of living during the Industrial Revolution in Britain (for an exercise on this article see p. 232). It graphs the movements of a new food price index constructed by Feinstein with an older retail price index recorded by contemporaries for Oldham, Manchester and Staffordshire. The results are reassuringly similar although the fluctuations in the new index after 1820 are likely to be a more accurate reflection of food prices paid by the mass of the population than the older series.⁵

Choice of base year can be important in creating an impression of change in an index. If a low value near the start of the series is chosen, the index may appear at first glance to be growing much more significantly than if a later, higher figure was chosen. It is normally best to choose a base year near the middle of a series and a year that is not markedly out of line with the rest of the values or any perceivable trend. There may be other good reasons for the selection as in Table 5.3 where 1700 has been selected largely because the observations cover only England before this date but Great Britain thereafter. Having the base year 1700 enables the figures both before and after 1700 to be viewed separately without the potentially distorting effect of choosing a base year either earlier or later.

From the data in Table 5.1 1866 was chosen as the base year because it is near the middle of the series and the value for that year is not markedly out of line (unlike 1865). In Table 5.2 1929 was chosen as the base year for both series for the same reasons. Feinstein (Figure 5.2) chose 1791 because it was the first year for which both series were available but also because it is not out of line with other readings compared with values in the period 1798–1820.

History by Numbers

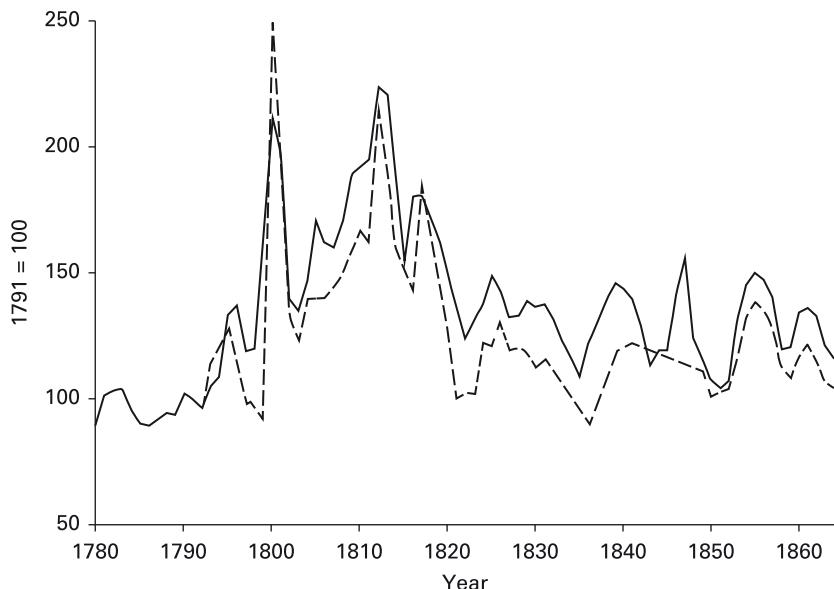


Figure 5.2 Indices of food prices, 1780–1870. Solid line = Feinstein's food index; dashed line = retail price series.

Source: C. H. Feinstein, 'Pessimism perpetuated: real wages and the standard of living in Britain during and after the industrial revolution', *Journal of Economic History*, 58, 3 (1998), p. 637.

Once indices have been formed it is possible to go on to produce composite and real indices.

Composite indices

A **composite index** is an index combining the simultaneous movement of several variables in weighted combination.

Composite indices are commonly used in estimates of the movement of average wages which are based upon figures for different occupational groups. They are also used a lot in estimates of the movement of the cost of living based upon price variation in the major components of family expenditure. Estimates of industrial output based on output figures from key sectors and estimates of change in agricultural prices based upon price series for individual crops and other products are further examples of the sort of information that composite indices can be formed to calculate.

To form a composite index one must follow four steps:

Step 1

First it is necessary to decide which series to include in the composite. This is an especially important decision to be made when estimating the movement of living costs because

there may be many different price series that could be included. It will be necessary first to decide what the main components of living costs were likely to have been. To do this a so-called **basket of goods** is compiled with the major components represented. Price series evidence is then sought which will reflect the price movements of these components. Similarly if constructing a composite index of the output of the economy based upon the output of many different sectors, the most important and representative sectors must be selected and no significant ones should be entirely absent unless their growth can be proxied by some other variable that is easier to obtain. For example population growth is often used as a proxy for the growth of the service sector.

Step 2

All of the separate indices which are to be included in the composite must have the *same base year*. If they do not already it is very easy to convert a series to a new base year as shown in Table 5.4 and explained below.

Step 3

The next step is to make (often-difficult) decisions about **weights** based on your judgement as to the relative importance of each individual series in the overall index. If for example one is estimating an index of the movement of living costs you will need to decide what proportions of the ‘basket of goods’ comprise rent, fuel, clothing, food, transport and so on, so that you can give the changing prices of these items their due emphasis in the overall composite index (see Allen’s example in Figure 5.3). Similarly if one is estimating the movement of Gross Domestic Product of an economy based on the output figures of various sectors, it will be necessary to decide what weight to give each sector in the overall calculation. Sometimes this is done using the national workforce distribution as a guide. There are some good examples of the calculation of weighted composite indices in Tables 5, 6 and 7.

Table 5.4 Indices of wages of industrial and agricultural workers (selected years)

Year	Indices of wages		
	agricultural 1890 = 100	industrial 1880 = 100	recalculated industrial 1890 = 100
1870	89	93	90
1875	88	95	92
1880	92	100	97
1885	96	98	95
1890	100	103	100
1900	98	99	96
1905	101	108	105
1910	99	110	107

Source: Hypothetical data.

History by Numbers

Step 4

Finally, one must multiply each index number by its weight, add these together and then divide by the sum of the weights.

The common base year

Before giving an example of the construction of a composite index it is necessary to demonstrate how to convert indices to a common base year. Table 5.4 provides a simple bare bones example using hypothetical data. It gives index numbers for the wages of agricultural and industrial workers in the late nineteenth century. Before attempting to create a composite index it will be necessary to ensure that both series have the same base year. We here choose to convert the industrial wage series to base year 1890 to conform to the agricultural wage series. Conversion to the same base year enables much easier comparison between the two series showing the industrial wage series to be more buoyant than the agricultural wage series. To convert the industrial wage series to the base year 1890 each value of the old index was placed successively over the value for 1890 and multiplied by 100. For example, the new index number for the year 1870, $I^{(1870)}$, may be calculated as follows:

$$\begin{aligned} I^{(1870)} &= \frac{\text{index of 1870 in existing index (1880 = 100)}}{\text{index of new base year (1890) in existing index (1880 = 100)}} \times 100 \\ &= \frac{93}{103} \times 100 \\ &= 90 \end{aligned}$$

Similarly for 1905:

$$\begin{aligned} I^{(1905)} &= \frac{108}{103} \times 100 \\ &= 105 \end{aligned}$$

Construction of composite indices: some examples

In the example in Table 5.5 the wage indices of four groups of workers in England and Wales are given for the sample years 1780–1830. This first example uses hypothetical data to make the calculations as simple as possible. Once we have checked that they all have the same base year, to construct a composite index we must now give each component a weight. With an index of this kind we would need to make the best estimate which we can (based on complementary historical evidence) of the balance of each group of workers in the working population overall. We would need to justify our decision about the weights and point to the evidence which we have used. Weights can be expressed in any numbers which give an indication of proportion in relation to the whole. They are normally expressed in numbers which add up to 10 or to 100 which

Table 5.5 Indices of average money earnings per week (1890=100) and the formation of a composite index for selected years in the period 1780–1830

Year	Index				Composite index
	agric.	skilled manu.	unskilled	service sector	
1780	65	69	55	53	59
1810	68	71	50	50	58
1815	72	73	64	52	65
1820	70	74	64	54	65
1830	73	78	68	58	69
Weight	3	1	4	2	

Note: agric. = agriculture; manu. = manufacturing.

Source: Hypothetical data.

makes calculation easier, but they do not have to add up to such round sums. In our example we might decide to assign weights as follows:

- Agricultural 3;
- Skilled manufacture 1;
- Unskilled workers 4;
- Service sector 2.

To form our composite we now take each index number in turn and multiply it by its weight. We then add these and divide by the sum of the weights. Thus the composite for 1780 was formed as follows:

$$\begin{aligned} Ic^{(1780)} &= \frac{(65 \times 3) + (69 \times 1) + (55 \times 4) + (53 \times 2)}{(3+1+4+2)} \\ &= \frac{590}{10} = 59 \end{aligned}$$

The composite for 1820 was formed from:

$$Ic^{(1820)} = \frac{(70 \times 3) + (74 \times 1) + (64 \times 4) + (54 \times 2)}{10} = 65$$

The main problem in forming an accurate composite index is getting the weights right. This is not a statistical issue at all but a matter of historical assessment and judgement. Weights are difficult to establish partly because of unreliable or partial evidence but also because weights usually change over time. The wrong weights can considerably distort the composite index and make it meaningless. So central is this difficulty that it is referred to as '*the* index number problem'.

History by Numbers

In our example above the bare figures disguise the many pitfalls in constructing wage indices even just for one sector. Obtaining accurate money wage data is difficult when few wage books survive and when many were paid in family groups, by piece rates or in kind. There were also very significant regional and local wage differentials and variations between different occupations within each sector to say nothing of the need to make allowance for female as well as male wages and for seasonal and more chronic unemployment and underemployment. From the patchy data available each sectoral wage series is itself inevitably already a composite with the many problems that the construction of that composite has entailed.

Much debate amongst economic historians occurs over weighting decisions, 'the index number problem'. In revisions to indices of industrial output in Britain in the eighteenth and nineteenth centuries, for example, debates have hinged upon two things: choice of which industries to include (which in turn has related to the existence and reliability of the data) and what weights to assign to each. The choice of the weights radically affects the resulting calculation of the rate of growth of industrial output overall. One of the earliest indices of British industrial output since 1750 was compiled by Walther Hoffman in 1939 and this became a foundation for much theorizing about the Industrial Revolution. In 1982, C. Knick Harley argued that Hoffman had overweighted the faster growing industries, especially cotton for the period 1770–1815.⁶ Harley's revision suggested 40 per cent lower output growth for industry in the classic Industrial Revolution period and has ushered in whole series of studies of slow growth and gradualism. Since 1982 further more minor revisions have been suggested by Crafts and Harley and by Jackson based on differing weights and compositions of the index and there have been other challenges based upon revisions to the price data used. The most recent contribution to these sorts of estimations is by Broadberry et al. in their study of economic growth in Britain over six centuries. Their results for the industrial output component of economic growth are summarized in Table 5.6.

As in Table 5.3, and for the same reason, the base year for all indices in Table 5.6 is 1700. 'Metals and mining', 'Textiles and leather', 'Other industries' are all themselves composite indexes of different elements of each industrial sector in weighted combination. The authors justify their choices by suggesting that the various output series that they use are representative of wider experience of the various industries for which there is no direct evidence. The weightings used to estimate the various composites and the final column (total industrial output) are fully explained in a three-page appendix.

Most historians choose to weight according to the money value of the output. This is measured by multiplying unit prices by the quantities purchased in some base year (hence the importance of collecting accurate price data, which is difficult). Use of the estimated labour force share for each industry or sector is an alternative. The choice of year upon which to base the weight is also a problem because of changing weights over time. If values in 1850 are used as weights for British industrial output in the nineteenth century, cotton will loom very large. If the weights are based upon 1913 values (after many years of much slower growth in cotton output than in other sectors), the role of cotton will be much smaller. There is no objectively correct way of adding up the elements

Table 5.6 Output of key industrial sectors, England, 1270–1700 and Great Britain, 1700–1870 (1700=100)

<i>Decade</i>	<i>Metals and mining</i>	<i>Textiles and leather</i>	<i>Other industries</i>	<i>Total industry</i>
1270s	ND	35	11	27
1300s	16	32	12	27
1350s	9	40	6	20
1400s	30	32	5	20
1450s	17	33	5	18
1500s	26	32	7	22
1550s	40	42	11	31
1600s	46	70	31	51
1650s	47	65	65	61
1700	100	100	100	100
1750s	156	151	105	132
1800s	555	298	193	271
1850s	3,638	1,337	566	1,163
1860s	5,052	1,436	802	1,480

Source: Stephen Broadberry, Bruce M. S. Campbell, Alexander Klein, Mark Overton and Bas van Leeuwen, *British Economic Growth 1270–1870* (Cambridge 2015), p. 139.

in a composite index especially where the weights are likely to change over the time period being studied. The choice must be made and justified according to historical judgement and with the purposes of the research in mind.

Various weighting schemes have been named after nineteenth-century investigators. A **Paasche index** uses estimates for the current or last year/time period as weights throughout the series. A **Laspeyres index** uses estimated weights for the initial year/time period. Some weighting systems use neither of these but attempt to change the weights over time usually by averaging some component of the base and current indices. This is a difficult practice not only because it is difficult to estimate when or at what pace weights may change but also because important variations in the indices may occur simply as a result of a sudden shift in the weights applied, especially if this is not phased in very gradually. These are the pitfalls to which the Broadberry et al. estimates are prone: they include 13 changes in the weightings between 1700 and 1870.

Very often composite indices are used to calculate change in the cost of living. The Retail Price Index measures the change from month to month in the average level of prices for the commodities and services purchased by nearly nine-tenths of the households of the UK. The index is based on a stratified random sample of households whose basket of goods and the weights attached to various goods is researched in some detail. (See Chapter 7 for discussion of sampling techniques including stratified and random sampling.)

Table 5.7 gives an example of the construction of a composite cost of living index. Apart from the problems of obtaining reliable price data for periods in the past, the main difficulty in establishing a cost of living index is getting the basket of goods right. Often

History by Numbers

Table 5.7 Components of an index of living costs, 1890–1900 (1900=100)

Year	Food	Rent	Clothing	Fuel	Sundries	Composite index
1890	101	93	102	80	89	97.68
1891	103	94	102	78	85	98.72
1892	104	95	101	78	81	99.20
1893	99	96	100	85	81	96.80
1894	95	96	99	73	75	93.08
1895	92	97	98	71	75	91.16
1896	92	98	99	72	75	91.52
1897	95	98	98	73	75	93.28
1898	99	99	97	73	74	95.68
1899	95	99	96	79	76	93.72
1900	100	100	100	100	100	100.00
Weight	60	16	12	8	4	

Source: A. L. Bowley, *Wages and Income in the United Kingdom Since 1860* (Cambridge 1937), pp. 120–121.

Reproduced in R. Floud, *Introduction to Quantitative Methods for Historians* (London 1973), p. 126.

the basket of goods is constrained by the data available or the use of unreliable proxy figures is encouraged if price data for an important item in the basket of goods is missing. One of the earliest cost of living indices for the nineteenth century was produced by N. J. Silberling. J. H. Clapham used the Silberling index in his estimates of the living standards of the working classes during the period of industrialization, but as T. S. Ashton pointed out, 'Silberling man' was a strange creature indeed:

He did not occupy a house, or at least was not called upon to pay rent. He allowed himself only a moderate amount of bread and very little porridge, and he never touched potatoes or strong drink. On the other hand, he got through quite considerable quantities of beef and mutton and showed a fondness for butter. Perhaps he was a diabetic. The ordinary Englishman of the eighteenth century would have been puzzled by him.⁷

Use of the inadequate Silberling index by Clapham left him open to attack from those whose interpretations of living standards during the Industrial Revolution were much more pessimistic.⁸

More recent research by Robert C. Allen and others attempts to compare living costs across continents and over long time periods. The main purpose of Allen's study is to try to distinguish at what point and why the trajectory of growth in Western Europe began to diverge markedly from both southern Europe and from Asia. Allen's argument revolves around the positive impact of having a relatively high wage economy calculated, for comparative purposes, as one that allows a certain level of 'respectable' rather than 'bare bones' subsistence, nutrition and choice about investment in education and skills for succeeding generations.

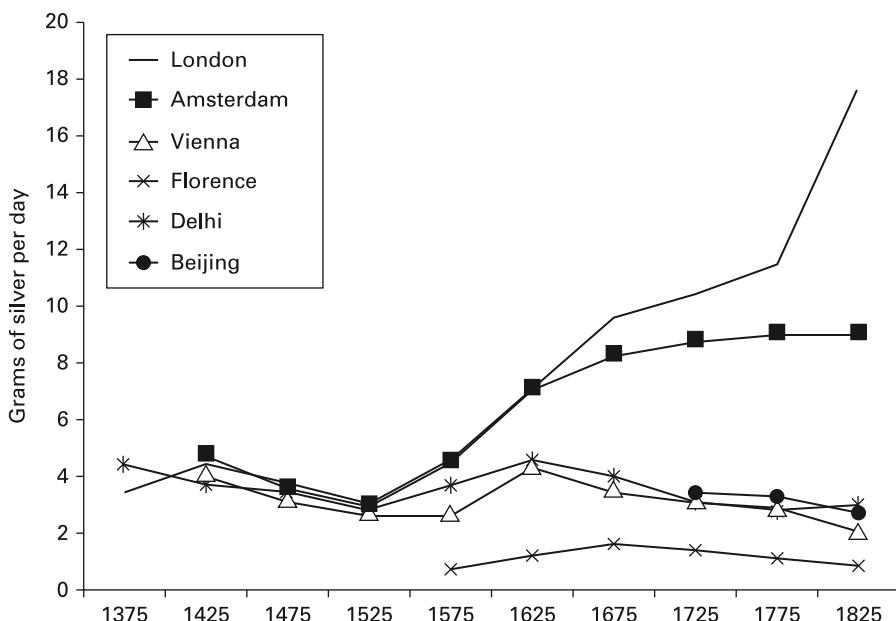


Figure 5.3 Labourers' wages around the world.

Source: Robert C. Allen, *The British Industrial Revolution in Historical Perspective* (Cambridge 2009), p. 34.

Figure 5.3 shows the results of some of Allen's work demonstrating the movement of wages in several locations in Europe and Asia. Enough has been explained already in this chapter to indicate the difficulties of calculating composite indices of wage movements even for just one country. With transnational comparisons of this kind not only is it difficult to get representative and comparable wage data but these have to be adjusted using a silver standard to allow for different currency values in the original data. A silver standard allows raw wage units in different currencies to be standardized according to the value of silver obtained internationally in the various currencies at each point in time. It assumes that the value of silver can be accurately determined and is fairly uniform across a country and a currency (a rather big ask).

Real indices

A **real index** is the movement of a time series, in index form, which has been adjusted to allow for the movement of another series normally prices. This gives a measure of real change (with the effect of deflation or inflation taken into account), for example, with real wages or real incomes or the values of exports or imports in constant prices.

To form a real index:

Step 1

You must have the *same base year* for each component index.

History by Numbers

Step 2

Divide the series to be adjusted (in the example below, the wage index) by the second series (in this example, the cost of living index) and multiply by 100 for each cell of information.

In Table 5.8 composite earnings estimates as calculated on p. 137 above have been matched with a price series to produce a real index of 'living standards' (insofar as these are indicated by changes in the purchasing power of wages alone). In Table 5.8 the two series did not originally have the same base year. The base year of the wage series was changed as indicated.

A further example of real indices is provided in Table 5.9 which shows gate receipts and admission prices at English Football League grounds in 1927–1994. Both gate receipts and admission prices are given in the original units of measure (£) but they have also been converted into indices and deflated using the Retail Price Index. Indices make the figures easier to compare with one another and the real indices convey change in the real cost of attendance at games for the fans and the purchasing power of revenues received by the clubs.

Figure 5.4, derived from Allen's work, shows the movement of living standards in different locations in Europe and Asia over five centuries, at fifty-year intervals. At each benchmark year the average income of a representative sample of working people is adjusted with a cost of living metric that is seen as providing similar levels of food (calorific values) and fuel, with a margin above basic subsistence. In other words the respectability ratio is a **real index** composed of two composite indices that provide the numerator and the denominator. Problems with the wages series in transnational comparisons of this kind were raised above. Here, in addition, the composite cost of living index has to be standardized across societies that have different food staples, nutritional and social needs.

Table 5.8 Construction of an index of real wages, 1890–1900

Year	Money wages (1914 = 100)	Money wages (1900 = 100)	Cost of living (1900 = 100)	Real wages (1900 = 100)
1890	83	88.3	97.7	90.4
1891	83	88.3	98.7	89.5
1892	83	88.3	99.2	89.0
1893	83	88.3	96.8	91.2
1894	83	88.3	93.1	94.8
1895	83	88.3	91.2	96.8
1896	83	88.3	91.5	96.5
1897	84	89.4	93.3	95.8
1898	87	92.6	95.7	96.8
1899	89	94.7	93.7	101.1
1900	94	100.0	100.0	100.0

Source: Based on R. Floud, *Introduction to Quantitative Methods for Historians* (London 1973), p. 123.

Original source: money wage index from E. C. Ramsbottom, reprinted in B. R. Mitchell and P. Deane, *Abstract of British Historical Statistics* (Cambridge 1962), p. 345. Cost of living index is taken from Table 5.7.

Table 5.9 Aggregate league attendance, gate receipts and average admission prices, 1927–1994

Year	Attendance ^a		Gate revenues ^b		Real gate revenue (1927 = 100)	Average admission price (£)	Real average admission price (1927 = 100)
	total	coef. var.	total	coef. var.			
1927	23.4	0.65	1373	0.69	100	0.06	100
1932	21.8	0.66	1263	0.73	110	0.06	118
1937	26.4	0.63	1575	0.72	133	0.06	118
1947	35.4	0.61	2933	0.70	183	0.08	121
1952	39.0	0.57	4135	0.65	197	0.11	118
1957	32.7	0.57	4311	0.67	171	0.13	122
1962	28.0	0.63	4981	0.76	175	0.18	146
1967	28.9	0.73	6931	0.90	205	0.24	166
1972	28.7	0.77	10814	0.95	238	0.38	194
1973	25.4	0.85	11823	1.00	241	0.46	222
1974	25.0	0.79	13174	0.97	239	0.53	223
1975	25.6	0.81	15180	0.98	228	0.59	209
1976	24.9	0.83	18822	0.97	231	0.76	217
1977	26.0	0.82	22220	0.97	234	0.85	210
1978	25.4	0.84	26651	1.00	257	1.05	236
1979	24.5	0.83	28960	0.95	254	1.18	243
1980	24.6	0.79	36911	0.95	272	1.50	258
1981	21.9	0.82	40239	1.02	264	1.84	281
1982	20.0	0.84	40523	1.04	239	2.03	279
1983	18.8	0.82	42096	1.03	236	2.24	294
1984	18.3	0.82	44760	1.06	239	2.44	304
1985	17.8	0.92	49276	1.17	249	2.77	327
1986	16.5	0.94	48901	1.15	236	2.97	334
1987	17.4	0.89	55844	1.08	259	3.21	348
1988	18.0	0.81	63906	1.00	287	3.56	373
1989	18.5	0.77	72885	0.98	304	3.95	384
1990	19.5	0.77	87219	0.97	337	4.48	405
1991	19.5	0.81	103691	1.04	369	5.32	442
1992	20.4	0.83	127329	1.09	435	6.25	499
1993	20.6	0.77	146238	1.10	485	7.09	549
1994	21.7	0.82	163655	1.10	534	7.55	576

^aMillions. ^b£Thousands.

Note: In order to keep the table to a manageable size, figures are given for every fifth year only, up to the early 1970s. Years are end-years of football seasons; i.e. 1927 is the 1926–7 season, and so on. Gate revenues and admission prices are deflated using the Retail Price Index. Coef. var. = coefficient of variation.

Source: S. Dobson and J. Goddard, 'Performance, revenue and cross subsidization in the Football League, 1927–1994', *Economic History Review*, 51, 4 (1998), p. 767.

History by Numbers

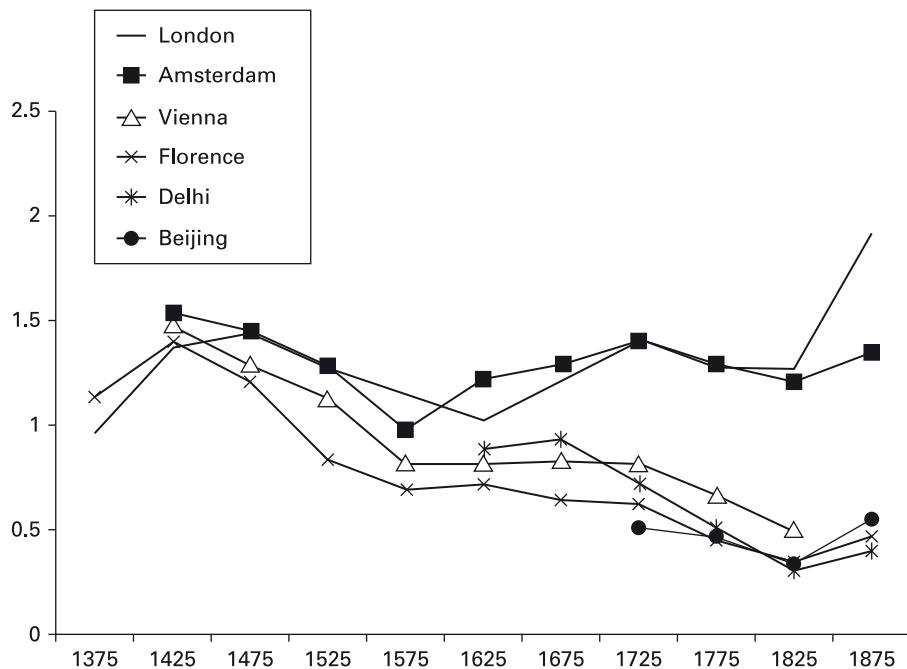


Figure 5.4 Respectability ratio for labourers: income/cost of respectable basket.

Source: Robert C. Allen, *The British Industrial Revolution in Historical Perspective* (Cambridge 2009) p. 39.

The calculations behind these figures are thus fraught with error but they are thoroughly explained and justified in the wider research that Allen draws upon.⁹

Time series: influences

There are several methods of statistical analysis of time series. Most centre on the problem of separating out or isolating the various sorts of change and the component causes of change in the value of a variable over time. The methods of time series analysis assume that there may be three types of influence affecting any time series:

1. *Trend influences* affecting long-term growth or decline.
2. *Regular fluctuations* around the long-term trend, caused by seasonal or cyclical factors.
3. *Irregular fluctuations*: short term, generally unrepeated movements caused by, for example, wars, diseases, changes in government policy.

We must use historical judgement to ask if any or all of these three influences may be present in a series before we attempt to isolate and examine each. It is wise to graph the series to get an idea of its character before undertaking any more complex time series analysis. In a time series graph it is conventional to place time on the horizontal axis. In

Figure 5.5 all three influences appear to be present. There is an interrupted but perceptible upward trend in burials over time, there are fluctuations that appear to have some regularity, and there are individual years of exceptionally high or low levels of death. The baptism series has a less obvious upward thrust until the late 1570s and, again, there are some years of exceptionally high or low levels even allowing for the possibility of defective registration. (Defective registration can be caused by a number of things such as illness or death of the incumbent clergyman, periods of war or social disturbance, lost pages in the register, illegibility of the register.)

TRENDS

There are several ways in which a **trend** may be identified and measured. As mentioned above, the simplest way to approach the nature of the time series initially is to graph it. This alone may highlight the trend sufficiently for certain analytical purposes. If the time series variable on the graph is generally upward sloping from left to right this indicates a *positive trend* (growth over time). If the time series variable is generally downward sloping this indicates a *negative trend* (decline over time). The steeper the slope the greater the rate of growth or decline.

Not all trends in data are linear (that is, tending towards a straight line trajectory in one direction). Sometimes trends are present in data which are non-linear: the observations do not lie around a straight line but around a curve of some sort. This is the

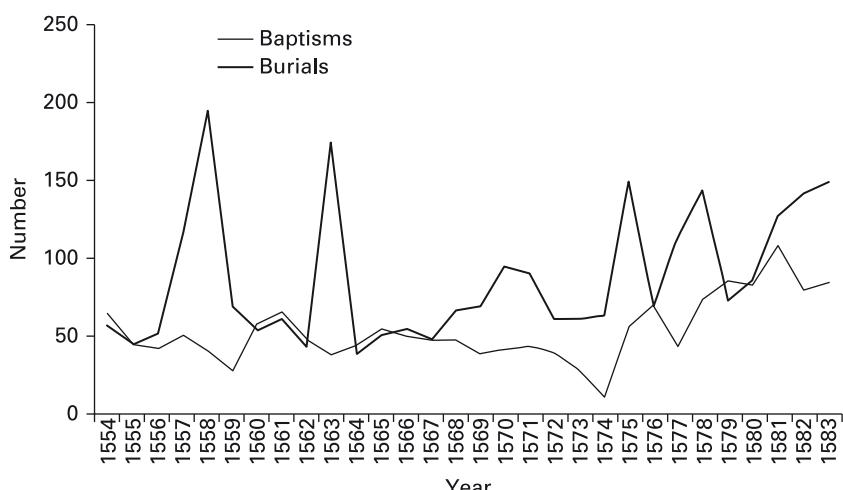


Figure 5.5 Time series graph of burials and baptisms, St Martin in the Fields, London, 1554–1583.

Source: Parish registers, St Martin in the Fields, London, 1554–1583.

History by Numbers

case for example with various measures of inequality in incomes and in wealth holding in developed economies between the late nineteenth century and the present. These cannot be analysed in the manner described below and it is always wise roughly to graph any time series data before engaging in more sophisticated calculations, in order to check that a linear trend may be present. It is possible to analyse time series that embody non-linear trends but this is less commonly undertaken and beyond the scope of this volume. Graphing the data at the outset may also highlight a shift in the slope of the trend (reflecting change in the growth rate) at a particular point in time: this suggests a need to measure growth rates for certain subperiods within the data as well as over the series as a whole.

Measures of trend: growth rates

Growth rates are often used as a general measure of the pace and direction of trend in a time series. Growth rates can be positive (when the values of a series are generally increasing over time) or negative (when the values of a series are decreasing over time). There are several ways to measure growth rates but with all of them it is important to acknowledge at the outset that the choice of period over which the growth rate is to be measured can make a big difference. Many time series exhibit marked cyclical fluctuations. This is particularly true of national income and its components and determinants (for example, Gross Domestic Product, industrial output, exports). If we choose to measure growth rates from the depth of a slump at the start of the series to a major boom at the end, the growth rate is likely to be seriously inflated. If, conversely we choose to measure from a boom year at the beginning to a slump year at the end the growth rate is likely seriously to underestimate growth in the period. Both should be avoided.

Two different growth rates are commonly used: the *mean increase per year* and the *average percentage growth rate*.

The *mean increase per year* is calculated by using the formula:

$$\text{mean increase per year} = \frac{X_N - X_T}{N}$$

Where

X_T is the value of the variable at the start of the series;

X_N is the value of the variable at the end of the series;

N is the number of observations.

If the result of the calculation is negative a rate of decline rather than growth is indicated.

Table 5.10 and Figure 5.5 show a positive trend of burials, a less obvious one for baptisms and, as Table 5.10 suggests, marriage numbers, though fluctuating, seem relatively stable over the length of the series. The mean increase per year for burials, taking the first and the last observation, gives a result of 3.13 burials ((150–56)/30).

This result probably exaggerates the underlying growth rate because the last observation is one of a group of particularly high numbers of burials. The growth rate (as measured by the mean increase per year) for marriages is 0 because the higher

Table 5.10 Annual baptisms, marriages and burials, St Martin in the Fields, London, 1554–1583

Year	Number of baptisms	Number of marriages	Number of burials	Plague death ^a
1554	64	25	56	
1555	44	15	45	
1556	42	12	52	
1557	50	24	115	3
1558	40	18	195	1
1559	27	23	69	
1560	58	25	54	
1561	65	16	61	
1562	48	19	43	
1563	38	23	175	145
1564	44	25	38	1
1565	55	29	51	
1566	50	28	55	
1567	47	22	48	
1568	47	23	66	
1569	39	28	69	
1570	41	23	94	
1571	43	29	90	
1572	39	26	61	
1573	28*	18*	61	
1574	11*	6*	63	
1575	56	29	150	51
1576	70	35	69	4
1577	43*	25	114	14
1578	74	41	144	
1579	86	39	73	
1580	83	26	86	
1581	109	30	128	
1582	80	30	142	
1583	84	25	150	

^aNumber of deaths said to be the result of the bubonic plague.

*Defective registration.

Source: Parish registers, St Martin in the Fields, London, 1554–1583.

History by Numbers

levels of the 1570s are not reflected in a measure that relies only on the first and last observation.

This is the main problem with using the mean increase per year: the values of the first and last observation become all-important. This is a serious problem unless the series is growing or declining very steadily and without major fluctuations. If the trend is unstable the first or last reading may be markedly out of line with the general trend and will distort the growth rate result. If, on graphing the data, it is obvious that the first or last reading are widely out of line it is wise to take another more representative value near to the start or end of the series instead. If this is done the value of N will need to be adjusted accordingly.

In the example given (Table 5.10, Figure 5.5) the first reading for baptisms is markedly out of line and it would give a more accurate measure of the growth rate for the whole series if we took the second rather than the first observation. If we were to draw two lines on the baptism graph in Figure 5.5 from the first to the last observation and from the second to the last, the two lines with their different gradients (and hence different rates of growth) illustrate the difference made by choosing a start date that is markedly out of line compared with a more representative observation. Taking the second observation the mean increase per year becomes 1.3 baptisms rather than 0.7 baptisms per year.

It is worth noting the following:

- (i) The unit of measurement in which the mean increase is expressed is the original unit of the series (for example, number of marriages or burials, exports, crimes, incidents of industrial sabotage) it is thus impossible to compare one growth rate with the growth rate of another series, expressed in different units. We may, for example, wish to compare the growth rate of export values over time with the tonnage of the merchant fleet, the number of people able to sign marriage registers with investment levels in education or incidents of industrial sabotage with wage rates. For these sorts of **growth rate comparisons**, a different measure is needed. One possibility is the *average percentage growth rate*.
- (ii) The *average percentage growth rate* is not simply the average of the growth rates from year to year in a series. Such a measure would overestimate growth because growth is cumulative. What is needed is a measure that expresses each year's growth as a percentage of the value of the previous year. This is the purpose of the average percentage growth rate: it measures the average of the increase of each year or period over the previous one (in other words, the compound growth rate), expressed as a percentage. It eliminates the problem of needing a common unit for comparison (as the growth rate is expressed as a percentage) but it still suffers from reliance upon only the first and last observations chosen.

The formula for the *average percentage growth rate* (r) is as follows:

$$r = \left[m \sqrt{\left(\frac{X_N}{X_T} \right)} - 1 \right] \times 100$$

This is exactly the same as:

$$r = \left[\left(\frac{X_N}{X_T} \right)^{\frac{1}{m}} - 1 \right] \times 100$$

Where

r = the average percentage growth rate;

m = the difference in years between the first and the last reading;

X_T = the value of the variable at the start of the series;

X_N = the value of the variable at the last reading.

A computer package will make the calculation automatic once the key figures are inputted.

If we wish to compare the growth rates of the two vectors, baptisms and burials [$r(\text{baptisms})$ and $r(\text{burials})$] respectively, from Table 5.10 we could use the following calculation:

$$r(\text{baptisms}) = \left[\left(\frac{84}{44} \right)^{\frac{1}{28}} - 1 \right] \times 100 = 2.3$$

$$r(\text{burials}) = \left[\left(\frac{150}{56} \right)^{\frac{1}{29}} - 1 \right] \times 100 = 3.5$$

The baptism growth rate employs the second and the last observation to avoid 1554 which is anomalous. The growth rate of burials is greater than the growth rate of baptisms even though the latter is probably exaggerated more than the former by the high value of the last reading. This suggests that urban mortality rates in the sixteenth century were high and that parishes such as St Martins could only maintain or increase their populations through immigration.

Calculation of the trend line

The **trend** is an alternative and often better measure of the pace of change than growth rates because it takes account of all readings, not just the first and the last. It is formed by calculating and drawing the **line of best fit** through the series.

We can draw a trend line roughly through a series, after graphing the points as in Figure 5.6, but to get its position exactly right the sum of all the distances of observations above the line must equal those below.

History by Numbers

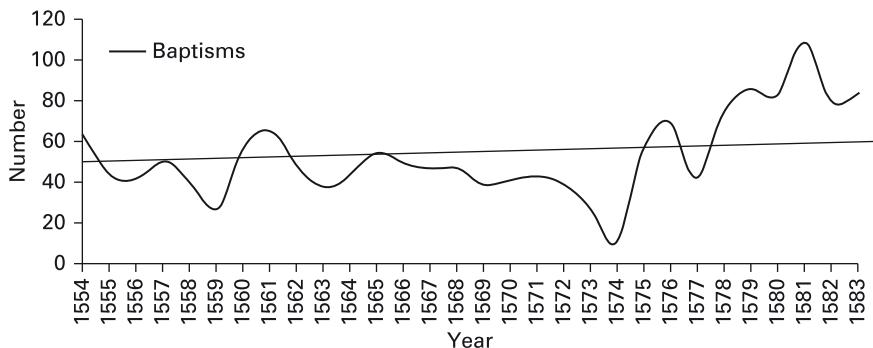


Figure 5.6 Trend line of baptisms, St Martin in the Fields, London, 1554–1583.

Source: Parish registers, St Martin in the Fields, London, 1554–1583.

We can calculate the exact position of the **trend line** by using the general equation for a straight line:

$$Y = a + bX$$

Where

Y is the data value;

X is the time unit;

a is the intercept (the value of Y when $X = 0$);

b is the slope.

The intercept a and the slope b of the trend line can be calculated as follows:

$$a = \frac{\sum Y}{N}$$

$$b = \frac{\sum XY}{\sum X^2}$$

Where

Y = data values;

N = number of observations;

X = time unit expressed so that $X = 0$ in the middle of the series.

We have here taken the time unit expressed so that X equals 0 in the middle of the series because it makes the calculation much simpler when it is done manually (in particular it makes the calculation of X and XY less unwieldy, as shown in Table 5.11). Such calculations are normally done using computer software, leaving the time units as in the original. In this case the formulae in use are:

$$a = \frac{\Sigma Y - b\Sigma X}{N}$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

Once we have the trend line equation we can find all the values of Y that lie on the trend by substituting values of X in the equation.

There are three things to note:

1. We can now get a much more accurate measure of growth rates by using the first and last value of the trend line instead of the first and last value of the original series (remember that the trend values have taken into account all the values of the series).
2. Once we have the linear trend we can subtract the trend value for each time period from the original data value. This will leave us with the **detrended series** composed only of any regular and irregular fluctuations that are present. We can now see these more clearly as a preliminary to further possible analysis.
3. We can use the trend line to forecast or to predict what the values of a variable might be in time periods later than that for which data are available, simply by extending the trend line. The possibility of **extrapolation** has many uses in history, in economics and in many other subjects. It has great potential in terms of the insights that can be gained but it must be remembered that useful forecasting is dependent upon both the accuracy in measuring existing values and upon the unchanging nature of influences that might affect the behaviour of a dependent variable more in one time period than another. The interpretation of extrapolations, as well as the decision to make extrapolations, is in the hands of the historian.

Following the example in Table 5.10 and Figure 5.6 (the baptism series), we can now create the trend values of baptisms and, by taking each of these trend values from the corresponding data value in turn, we can create the **detrended series** of baptisms as shown in Table 5.11. Spreadsheet or statistical software can make short work of this sort of calculation even for very long time periods, once the data have been inputted.

The formula for the trend line of baptisms is $Y = a+bX$. With respect to the data in Table 5.11, we now know that:

$$a = \frac{\Sigma Y}{N} = \frac{1521}{29} = 52.4$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{2120}{2030} = 1.04$$

The value of b is small because the trend for baptisms is very weak. This may be partly because we have included the years of defective registration which almost certainly

History by Numbers

Table 5.11 Calculation of the trend and of the detrended series of baptisms, St Martin in the Fields, London, 1554–1582

Year	Number of baptisms (Y)	Number of time units from 1568 (X)	X ²	XY	Trend values	Detrended series
1554	64	-14	196	-896	37.83	26.17
1555	44	-13	169	-572	38.87	5.13
1556	42	-12	144	-504	39.92	2.08
1557	50	-11	121	-550	40.96	9.04
1558	40	-10	100	-400	42.00	-2.00
1559	27	-9	81	-243	43.05	-16.05
1560	58	-8	64	-464	44.09	1.91
1561	65	-7	49	-455	45.14	19.86
1562	48	-6	36	-288	46.18	1.82
1563	38	-5	25	-190	47.23	-9.23
1564	44	-4	16	-176	48.27	-4.27
1565	55	-3	9	-165	49.32	5.7
1566	50	-2	4	-100	50.36	-0.36
1567	47	-1	1	-47	51.40	-4.40
1568	47	0	0	0	52.45	-5.45
1569	39	1	1	39	53.49	-14.49
1570	41	2	4	82	54.54	-13.54
1571	43	3	9	129	55.58	-12.58
1572	39	4	16	156	56.63	-17.63
1573	28	5	25	140	57.67	-29.67
1574	11	6	36	66	58.71	-47.71
1575	56	7	49	392	59.76	-3.76
1576	70	8	64	560	60.80	9.20
1577	43	9	81	387	61.85	-18.85
1578	74	10	100	740	62.89	11.11
1579	86	11	121	946	63.94	22.06
1580	83	12	144	996	64.98	18.02
1581	109	13	169	1417	66.02	42.98
1582	80	14	196	1120	67.07	12.93
Total	1521 ^a		2030 ^b	2120 ^c		

^a ΣY ^b ΣX^2 ^c ΣXY

Source: Parish registers, St Martin in the Fields, London, 1554–1582.

significantly underestimate baptisms. We calculate the trend values from the formula and form the detrended series by subtracting each trend value from the corresponding data value in turn, as illustrated in Table 5.11.

Given the weakness of the trend in the baptism series, particularly before 1574, one might legitimately question the usefulness of calculating the trend here. The advantages of doing so are much more obvious where time series exhibit a more notable trend.

However, the trend line does allow a more accurate growth rate to be calculated, over the series as a whole and taking each observation into account. Also the detrended series assists in identifying years that are markedly out of line. With such a weak trend and defective data the temptation to extrapolate or predict values for baptisms after 1583 or before 1554 should be firmly rejected.

REGULAR FLUCTUATIONS

By no means do all time series have regular, periodic fluctuations, but many do. The most common regular fluctuations embodied in time series are **seasonal** and **cyclical**. Seasonal fluctuations are commonly found in temperate parts of the globe and especially in pre-industrial data as the rhythm of economic, social and demographic activity was very much underpinned by climatic variations and the agricultural calendar. Seasonal factors continue to influence activity in many areas of economic and social life to the present day. Cycles of booming output and employment followed by years of relative depression are also common in most industrial economies as trade and investment tend to build up in a wave of confidence and optimism to a point at which interest rates rise, markets become overstocked and business confidence again lapses. Longer cycles of activity may also be present in long-run output and investment series because economic activity is influenced by the clustering of innovations around new products and/or services which occur every seventy years or so.¹⁰

Regular cycles occur in all sorts of time series. For example retail sales can rise and fall in relation to the seasons of the fashion year; food prices may rise and fall in relation to the regular swings of the harvest year; employment levels in jobs affected by seasonal demands rise and fall in relation to peaks of activity at Christmas or Easter, crime figures sometimes rise and fall in relation to the business cycle and the shifts in unemployment and income levels which this creates; profit rates, interest rates and many other series often shadow the regular fluctuations of the business cycle.

Cyclical fluctuations and moving averages

Cyclical fluctuations occur at regular intervals in a time series. It is useful as a first step to check if cyclical fluctuations appear to be present by drawing a graph of the series. The cyclical fluctuations that show up when a time series is presented graphically can be identified and discussed by the historian as part of her analysis. If, in the judgement of the historian, cyclical fluctuations of an identifiable periodicity are present in the series, it is possible to eliminate them so that the trend and irregular fluctuations can be viewed more closely. The latter are often the main focus of interest for the historian. Cyclical fluctuations can be removed by using a **moving average**.

A moving average is calculated and applied as follows:

History by Numbers

Step 1

It is first necessary to make an informed judgement about the periodicity of the cycle. Does the graph suggest a cycle of seven years (a common business cycle length) or of nine years (sometimes seen in demographic statistics) or of any other clearly identifiable length?

Step 2

Next one must form a new series. The first observation of the new series is taken at the midpoint of the first cycle and is the mean of all observations in the first cycle. Succeeding observations are the means of successive cycle-long groups of observations. The formation of a moving average is illustrated in the example in Table 5.12.

The problems with applying a moving average are as follows:

- (a) Values for several years (half a cycle) at the beginning and end of a series are lost.
This is serious if a long cycle is present and if the data only covers a small number of complete cycles.
- (b) The method only works when we are fairly sure of the periodicity of the cycle. A wrong choice of cycle length (illustrated in Figure 5.7) can produce extremely misleading results which invert the appearance of cyclical change in the data.
This must be avoided at all costs so the use of a moving average to eliminate a regular cycle from a series must be done very carefully and only when there is some certainty that the correct periodicity of the cycle has been identified.

Smoothing the data in a long time series

Moving averages are often simply used to smooth the data in a graph of a very long time series so that trends and changes in trends can be examined. Choosing the wrong periodicity is not so crucial in this case though a correct periodization should always be the aim. A nine-year moving average is commonly used to smooth figures in demographic analysis, for example, because a decade or so normally includes short runs of bad harvests and similar economic conditions that can affect vital events totals. If cycles are not obvious but smoothing is desirable to get rid of exceptional observations, a five-year moving average is often selected, as in Table 5.12.

A further example of the application of a moving average is provided in Dunstall's study of violence in colonial New Zealand. In Figure 5.8 the homicide rate (number of homicides per 1,000 population) has been graphed for the period 1878–1980 using a three-year moving average to smooth out the data and to make the trend elements show up more clearly. Had the graph traced the data in its raw state the shifts from year to year between no murders and four or five murders would have introduced far too much variation (or 'noise') into the graph.

The graph as it stands shows a clear 'U-shaped curve' of declining and then rising violent crime in New Zealand over the last 150 years. The assault rate is included on the graph, this time without a moving average most likely because, as shown on the right-hand axis, the variation in assaults per 1,000 population had a much larger range and

Table 5.12 Mean heights (in inches) and the five-year moving average of mean heights (in inches) of English rural-born female convicts, aged 21–49 years, 1788–1819

<i>Year of birth</i>	<i>Number of convicts</i>	<i>Height</i>	<i>Five-year moving average</i>
1788	7	61.71	
1789	17	61.90	
1790	9	61.53	61.91
1791	15	62.35	61.66
1792	15	62.08	61.56
1793	16	60.47	61.51
1794	11	61.36	61.25
1795	15	61.28	61.17
1796	16	61.06	61.47
1797	18	61.69	61.59
1798	19	61.93	61.68
1799	12	61.98	61.89
1800	16	61.73	61.85
1801	29	62.11	61.67
1802	21	61.48	61.74
1803	29	61.50	61.64
1804	44	61.89	61.61
1805	30	61.24	61.58
1806	30	61.93	61.68
1807	37	61.35	61.65
1808	36	62.00	61.77
1809	33	61.74	61.72
1810	47	61.82	61.67
1811	37	61.68	61.53
1812	47	61.13	61.34
1813	29	61.29	61.19
1814	29	60.78	61.23
1815	28	61.06	61.05
1816	23	61.89	61.12
1817	6	60.21	61.48
1818	10	61.68	
1819	6	62.58	

Source: R. V. Jackson, 'The heights of rural-born English female convicts transported to New South Wales', *Economic History Review*, 59, 3 (1996), p. 586.

varied much less than homicides from one year to the next. The U-shape here is clearly visible without the use of a moving average. It is explained in the article that this trend mirrors that in countries of north-western Europe. The author highlights the close parallel between the colonial trend of violence and that in contemporary Britain and discusses the extent to which the colonial experience can be explained by frontier explanations and/or local cultural determinants. (See the exercise on pp. 119–120.)

History by Numbers

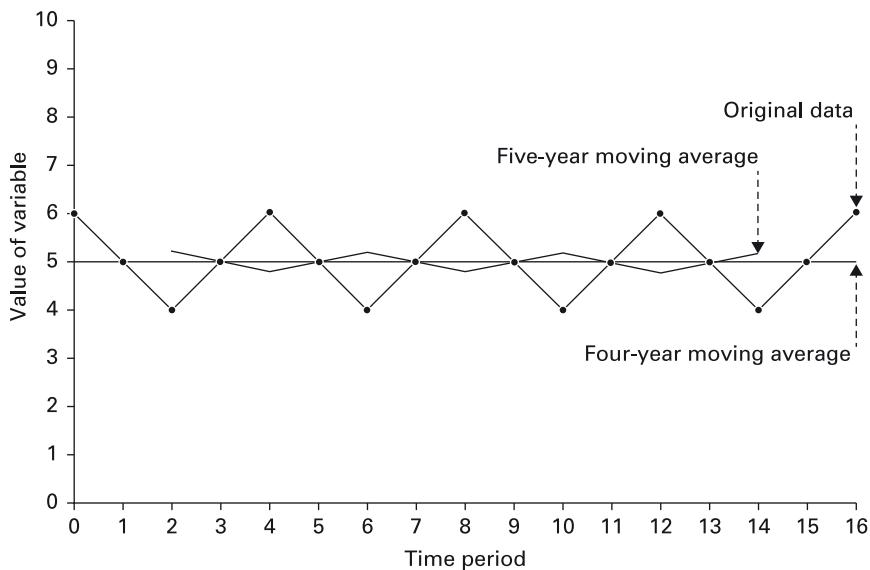


Figure 5.7 Stylized graph to show the impact of selecting a correct (four-year) and an incorrect (five-year) moving average.

Source: R. Floud, *Introduction to quantitative methods for historians* (London, 1973), p. 118.

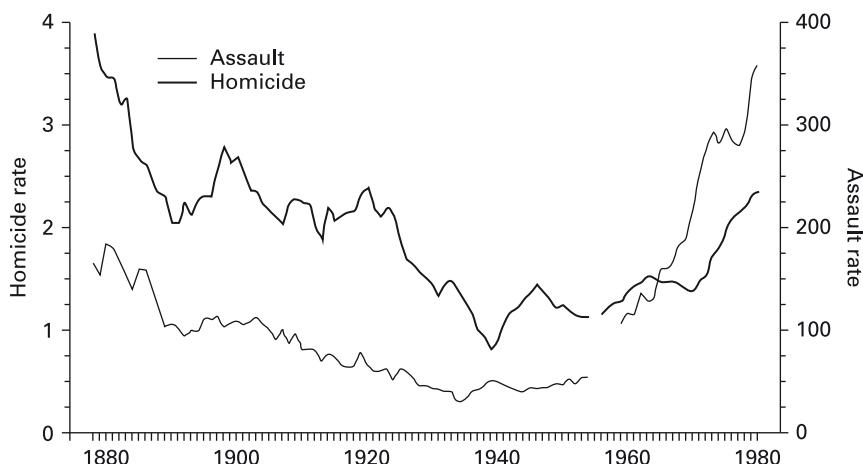


Figure 5.8 Reported homicides (three-year moving average) and reported assaults per 100,000, New Zealand, 1878–1980.

Source: G. Dunstall, 'Frontier and/or cultural fragment? Interpretations of violence in colonial New Zealand', *Social History*, 29 (1), (2004), p. 61.

Seasonal fluctuations

Seasonal fluctuations may be present in any time series with intervals of less than a year (for example, quarterly sales figures, monthly unemployment statistics, monthly or weekly food prices). They can be identified by taking the mean value of the variable concerned for each week, month, season or quarter and comparing them. If the original data contain an obvious trend, analysis of seasonality should be undertaken with the **detrended series** because trend elements will make the seasonal variations more difficult to observe.

Often it is useful to eliminate seasonal variations from a time series so that irregular fluctuations and/or trends can be viewed more easily. Most official statistics of economic activity, for example house prices, unemployment, agricultural output values, are 'seasonally adjusted' before they are published.

To separate seasonal fluctuations from a quarterly series:

Step 1

Calculate the trend values.

Step 2

Take the mean value of the deviations from trend at each quarter (or season, or week and so on), that is, the mean of all the first-, of all the second-, all the third- and all the fourth-quarter deviations.¹¹

Step 3

Subtract these means from the original values (the actual observations) at each quarter.

It is important to use the deviations from the trend series in step 2 rather than the raw data if a linear trend is present. If the trend series is not used the estimates of seasonal variation are distorted. The seasonally adjusted figures leave the trend easier to observe, as in Figure 5.9. The seasonally adjusted figures also leave the residual non-seasonal, one-off fluctuations easier to identify, as in Table 5.13 which uses hypothetical data to best illustrate the point. These residuals can then become the focus of enquiry if so desired. Table 5.13 gives the cost of provisions purchases in the Barrow workhouse in the 1880s. The trend values have been calculated as have the seasonally adjusted costs. The latter have been calculated by taking the mean deviation from trend for each quarter (-1.05, -7.8, 2.3 and 6.5, respectively) from the original series to create the new series of seasonally adjusted figures (rounded up to one decimal place). The seasonally adjusted costs have been graphed in Figure 5.9. These highlight the trend in the data free from seasonal bias.

In Table 5.13 the trend figures are not derived from the least squares linear calculation but employ a moving average method to avoid imposing a linear trend on the data (this is why there are no trend figures for the first and second quarters of 1883). The moving average method is often appropriate and preferable for establishing trend and deviation

History by Numbers

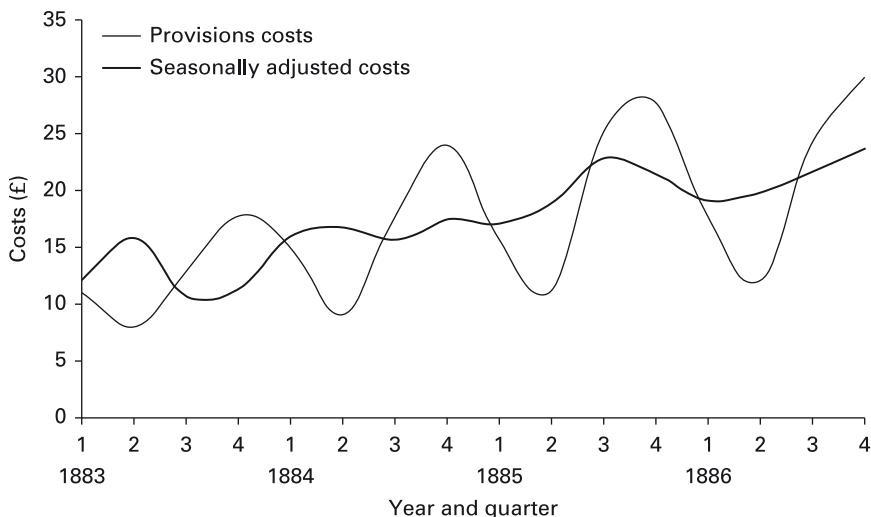


Figure 5.9 De-seasonalized movement of costs of provisions, Barrow workhouse, 1883–1886.

Source: Hypothetical data.

from trend, prior to de-seasonalizing data but the linear trend figures can also be used. There is a worked example of this in Table 6.16.

It should be noted that there are other more sophisticated methods of isolating seasonal variation, some of which are described in the texts in the further reading section at the end of this chapter.

If the periodicity of the cycle is known *with some certainty* it is possible to contemplate using the same technique to eliminate the cycle from the series as that used for getting rid of seasonal variations. If a seven-year cycle is present, for example, one could take the mean value of each year one reading, of each year two, three and so on, and subtract these from the actual readings of the detrended series. This would have the advantage over a moving average of highlighting even more clearly any irregular short-term fluctuations over and above those which one would expect from cyclical activity.

Figure 5.10 gives quarterly cash and credit sales of the Bradford wool stapling firm of Jowett and Co. The graph shows that credit sales were more important to Jowett and Co. than cash sales but that they were both subject to seasonal fluctuation. No trend is apparent from the graph so there would be no need to detrend the series before looking more closely at the regular and irregular fluctuations in this case. By taking the means of successive quarters from the actual readings for each quarter the seasonally adjusted figures for the credit sales could be established.¹²

Another example of the need for seasonally adjusted figures can be found in Odell and Weidenmier's study of the global financial impact of the San Francisco earthquake of 1906. In studying the gold outflows from London to San Francisco as a result of insurance claims following the quake, and in order to separate these out from normal

Table 5.13 Seasonally adjusted costs and residual fluctuations in provisions costs, Barrow workhouse, 1883–1886

Year and quarter	Provisions costs (£)	Trend (non-linear)	Seasonally adjusted costs (£)	Seasonal variation	Residual
1883					
1	11		12.1	-1.05	
2	8		15.8	-7.8	
3	13	13	10.7	2.3	-2.3
4	18	13.6	11.5	6.5	-2.1
1884					
1	15	14.3	16.1	-1.05	1.7
2	9	15.8	16.8	-7.8	1
3	18	16.6	15.7	2.3	-0.9
4	24	17	17.5	6.5	0.5
1885					
1	16	18.1	17.1	-1.05	-1.1
2	11	19.5	18.8	-7.8	-0.7
3	25	20.3	22.7	2.3	2.4
4	28	20.6	21.5	6.5	0.9
1886					
1	18	20.6	19.1	-1.05	-1.6
2	12	20.8	19.8	-7.8	-1
3	24		21.7	2.3	
4	30		23.5	6.5	

Source: Hypothetical data

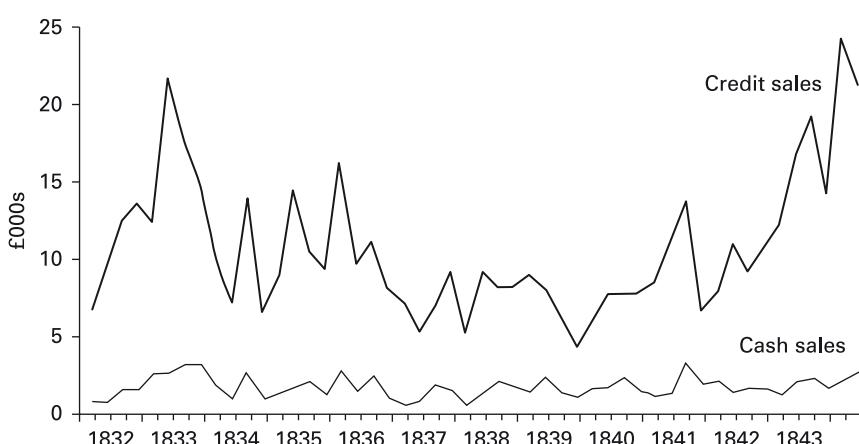


Figure 5.10 Quarterly cash and credit sales of Jowitt and Company, 1832–1843.

Source: Pat Hudson, *The Genesis of Industrial Capital* (Cambridge 1986), p. 207.

History by Numbers

seasonal flows of funds originating for other reasons, they provide seasonally adjusted figures that show that the \$11 million of gold imports arriving in San Francisco in late summer and early autumn of 1906 accounted for 9 per cent of all US gold imports.¹³ For a rather different purpose, Vincent Barnett uses seasonally adjusted figures of regional prices to consider the efficiency of supply and demand in Soviet commodity markets during the period of the New Economic Policy (NEP). He modifies seasonal variation to eliminate autocorrelation by using the series of first differences. This leaves the figures clear in demonstrating the degree of integration (or otherwise) of markets across vast swathes of the country. Seasonally adjusted figures for the USA are used as a comparator.¹⁴ Both of these articles are the subject of exercises on p. 227 and p. 228. Autocorrelation and first differences in the light of Barnett's research are explained on pp. 196–200.

Irregular fluctuations

It follows from the discussion above that if we wish to examine irregular fluctuations more closely, as Barnett does in his study of market integration, we may first detrend and/or seasonally adjust the time series. If a cycle is present and we can identify the periodicity, we can also eliminate cyclical influences by using the same technique or by using a moving average. This will leave us with a series of residuals containing irregular variations. We can then concentrate our energy on explaining the extent and the cause of these.

Table 5.13 shows the residual, irregular values that remain after detrending and deseasonalizing the time series of provisions costs. These are not large but the figures for the third quarters of 1883 (-2.3) and 1885 (+2.4) look sufficiently small and large respectively to warrant some explanation. Perhaps the number of inmates in the workhouse was the main cause of these variations. This could be checked if complementary sources, such as admissions registers were available.

In Barnett's comparison of the integration of Soviet and US markets, mentioned above, he shows that the raw data of many different agricultural commodity price series move in very similar patterns across regions but this says nothing about the integration of markets because most figures for food production incorporate major seasonal cycles. Only by eliminating such cycles can the degree of relationship between the price movements of specific commodities be compared across the Soviet Union or the USA for the purposes of deciding whether markets were well integrated or not.

Vital statistics or vital variables

Vital statistics are distributions relating to births, marriages and deaths. They are usually expressed in time series. Most often these are given as crude birth rates and crude death rates.

Crude birth and death rates are the total (live) births or deaths per 1,000 of the population during a period of one year:

$$\text{crude birth rate} = \frac{\text{total number of live births} \times 1000}{\text{total population}}$$

$$\text{crude death rate} = \frac{\text{total number of deaths} \times 1000}{\text{total population}}$$

If we wish to compare birth and death rates in two different populations separated by space or time, it is necessary to use standardized rather than crude vital rates. Standardized vital rates allow and adjust for differing age distributions of the populations concerned. They do this by weighting the crude birth or death rate by the age distribution of a standard population.

Conclusion

Time series are commonly used by historians who are frequently concerned with identifying and measuring the movement of different variables over time. There are many simple manipulations that assist in the measurement of growth or decline and that enable one to compare the movement of different variables over the same time period. Indices facilitate the comparison of variables over time and allow the construction of composite variables that incorporate the movement of several components in weighted combination. Real indices can also be produced. These reflect the movement of values which have been adjusted to take account of other variables (most often, price movements). The underlying trend in data can be calculated and then removed from the series, enabling regular and irregular fluctuations around the trend to be considered. Time series can be seasonally adjusted and moving averages can be used to eliminate cyclical fluctuations and to smooth irregularities so that trends can be more easily observed. In all, simple times series analysis offers a battery of techniques to the historian. Providing these are used with care, with sensitivity to the pitfalls, biases and inaccuracies of the original data, with due regard to the problems inherent in the construction of composite and real indices and with vigilance in the use of moving averages, they are an invaluable resource for researching the past.

That said, it is important to recognize that it is often the ordering hand of the historian that creates continuities and discontinuities in time series and establishes their timing:

Where we start and where we end and how we get there do not lie implicit and latent in the manner of history itself waiting only to be teased out by the skilled historian. Such matters are constructed by historians themselves as they order the material within certain categories and declare certain chronologies ‘periods’. In this process some things are suppressed, while others are privileged. It is sometimes thought that this allows historical statements only the status of fiction. Yet it is equally arguable that such artifices are enabling and powerful.¹⁵

History by Numbers

Further reading

Klein, Judy L., *Statistical Visions in Time: A History of Time Series Analysis, 1662–1938* (Cambridge 1997). (For those interested in how time series analysis has developed historically in statistical work.)

In most general statistics texts for social sciences time series are given little separate treatment but see:

- Darcy, R. and R. C. Rohrs, *A Guide to Quantitative History* (Westport 1995).
- Feinstein, Charles and Mark Thomas, *Making History Count: A Primer of Quantitative Methods for Historians* (Cambridge 2002), pp. 352–440, 446–460.
- Foster, Liam, Ian Diamond and Julie Jeffries, *Beginning Statistics. An Introduction for Social Scientists* (2nd edition, London 2015), pp. 47–50.
- Hanagan, T., *Mastering Statistics* (3rd edition, London 1997), Chapters 9 and 11.
- Haskins, L. and K. Jeffreys, *Understanding Quantitative History* (Cambridge, MA 1991), pp. 60–62, 289–291, 312–314.

CHAPTER 6

RELATIONSHIPS BETWEEN VARIABLES

A large circulation American Sunday newspaper published an article in the mid-1950s about the impact of a college education on the chances of getting married or remaining single. It reported that if you were a woman it hugely increased your chances of becoming an 'old maid'. For men a college education was found to have the opposite effect: it minimized your chances of remaining a bachelor. The underlying research from Cornell University had studied 1,500 middle-aged college graduates. Of the men 93 per cent were married (compared with 83 per cent in the population as a whole) but only 65 per cent of the female college graduates were married compared with almost 90 per cent in the general population.¹ Looked at closely, the article can be seen to be using a statistical association to support a possibly spurious cause-and-effect relationship. The figures could actually suggest not that college gets in the way of marrying but that women who are less disposed to marry are more likely to choose to go to college: in other words that college attracts feisty progressive girls who want their independence. College might modify these traits and if they had not attended it may be that even more would have failed to marry. Reading the research we would really like to know not just the size of the sample of graduates but how it was divided between men and women and how the respondents were selected. Was it justified to call them 'typical' and what 'middle-age' range was covered? We need to know the latter to understand which cohort of the population (that is, which social and age group) is being observed. Also, was the general population sampled in the same way to produce the contrasting figures or might we expect the national sample to reflect a different bias?² In addition, there is another common deception here: the statistics relate to Cornell graduates whilst the conclusions are generalized. Quantitative analysis frequently encourages the temptation to generalize from the particular to all cases especially where the analysis of relationships between variables is concerned. This chapter probes further the pitfalls as well as the advantages of quantitative research in history that concentrates upon the identification and force of causal relationships.

A question often asked by historians is: 'Does a relationship exist between two variables?' We might ask this about the relationship between educational levels and nuptiality rates, as above, or other cross-sectional data such as yearly income levels and size of mortgages of suburban dwellers. The same question might also be posed of variables in two or more time series (for example, the movement of exports and imports in the nineteenth century, or of average yearly grain prices and numbers of riots per year in the eighteenth century). Usually the question is provoked by some hypothesis about the causal connection which may lie behind a relationship (for example, education

History by Numbers

affecting nuptiality rates, income affecting mortgage size, exports helping to finance imports and food prices/hunger precipitating riots).³

Statistical techniques make it possible to investigate possible relationships with some precision and also to enquire into the strength and form of relationships. But, as we shall see, identifying and measuring relationships are not primarily matters of statistical technique. Statistical analysis can tell us *nothing* about the reliability of the data in the first place and can only indicate the *statistical probability* of a relationship being present. It cannot demonstrate the relationship itself. Identifying and assessing relationships between variables in a historical context involves historical judgement and common sense. It cannot be accomplished simply by using statistics. Statistics can only serve as a tool and can never substitute for historical analysis. The historian must choose when to apply statistical techniques and how to interpret the *significance* of statistical results.

It is important at the outset to form a hypothesis about the possible relationship between variables on the basis of sound historical judgement. Many sets of variables may by chance move or vary in a seemingly related fashion but this does not mean that there is any influence operating between the two or any causal connection. The British birth rate and the world stork population may well have similar variation over time, as may religious observance and the price of cabbages, but we would be foolish to suggest any meaningful connections. *Only if we can think of sound reasons why there might be a relationship between two or more variables should we indulge in the statistical identification and measurement of that relationship.*

The null hypothesis

Most statistical techniques concerned with the question ‘Is there a relationship between two variables?’ are based on a comparison between the dataset as it is and the dataset as it would be if there was no relationship at all. In other words, we pose our hypothesis that there is a relationship against the alternative **null hypothesis** that there is not. (The null hypothesis is conventionally expressed in shorthand as H_0 .) We then calculate the statistical probability of the existing distributions occurring by chance in the absence of a relationship. This gives us a measure of the strength of the relationship.

This way of testing a hypothesis, based on probability calculations and the null hypothesis was developed to deal with situations (as in most social science) where it is impossible to repeat experiments as a test. Being forced to work with the data that is available and the need to find some standard against which to measure the significance of observations gave rise to probability testing.

There are two commonly used techniques for testing for the possible existence and strength of a relationship between two variables and each involve use of the **null hypothesis**:

1. The contingency coefficient

This is suitable for nominal, ordinal, interval or ratio data.

2. The correlation coefficient

This is suitable only for interval and ratio data but is the most commonly used measure.

The correlation coefficient, as we mentioned in Chapter 2, was devised and promoted by Pearson who was committed to expressing all variables in the form of continuous numerical data. This arose from the demands of his eugenics research. His critics, including Yule, favoured methods akin to the contingency table where discrete nominal and ordinal data could be used.⁴

The contingency coefficient 'C'

The contingency coefficient is most frequently used in deciding whether a relationship exists between two variables that have been tabulated in the form of a **contingency table**. A contingency table is a particular form of a data matrix where two variables are plotted against one another. This is usually called a 'cross-tabulation'. The variables may be nominal, ordinal, interval or ratio data. In Table 6.1, sentence lengths (interval data) have been arranged in relation to type of crime (nominal data) using the information from Table 3.10.

There are one or two important points to note about the professional layout and presentation of contingency tables:

- All categories must be mutually exclusive.
- Each column must have a total and each row must have a total.

Table 6.1 Contingency table linking sentence lengths to types of crime, Portland Prison data, 1849

Length of sentence (years)	Type of crime					Total
	Unspecified felony	Larceny, housebreaking, burglary, stealing, theft ^a	Robbery ^b	Horse stealing	Forgery, incendiarism	
7–12	1	4	1			6
14–15		8	3	1	1	13
20				2		2
Total	1	12	4	1	3	21

^a Theft from property.

^b Theft from person.

Source: Based on Home Office, 8/102 The National Archives.

History by Numbers

- The total number of cases should be given in the bottom right cell of the table. This feature is important to aid the reader and as a check that you have included all cases.
- It almost goes without saying that all contingency tables, just like any other tables, should include a note or other reference indicating the source(s) of the data.

If we wish to enquire whether there is a relationship between the variables arranged in a contingency table it is first necessary to form an alternative contingency table based on the **null hypothesis**. This will give the expected frequencies (those which would occur in the absence of a relationship) to compare with the observed frequencies. In Table 6.2, level of education of men (ordinal data) has been related to family size (interval data) in a contingency table. The expected frequencies for level of education in relation to family size have also been calculated and the figures have been placed in brackets in the table. Hypothetical data is used here to make the calculation easier to understand.

In Table 6.2 the expected frequencies have been derived from probabilities based on the totals for each variable. For example, the probability of a man selected at random having a secondary education and 0 or 1 child is equal to the probability that he has a secondary education multiplied by the probability that he has 0 or 1 child. Such probabilities are of course unknown but they can be estimated from the sample data available. Thus the expected frequency for the above case E (secondary; 0 or 1), that is the value in brackets in cell A1,1 in the contingency Table 6.2 is calculated as follows:

$$\begin{aligned} E(\text{secondary}; 0 \text{ or } 1) &= p(\text{secondary}) \times p(0 \text{ or } 1) \times N \\ &= \frac{65}{86} \times \frac{32}{86} \times 86 \\ &= 24.19 \end{aligned}$$

Where:

p (secondary) is the probability of the man having secondary education;

p (0 or 1) is the probability of the man having no children or one child;

N is the total number of men in the sample.

Table 6.2 Contingency table linking level of education with family size (expected values given in parentheses)

Level of education	Number of children			Total
	0 or 1	2	3 or more	
Secondary	21 (24.19)	23 (21.16)	21 (19.65)	65
Higher	11 (7.81)	5 (6.84)	5 (6.35)	21
Total	32	28	26	86

Source: Hypothetical data.

Similarly, the probability (or expected frequency) of a man having a higher education and three or more children (cell A2,3 of Table 6.2) is:

$$E(\text{higher; } \geq 3) = \frac{21}{86} \times \frac{26}{86} \times 86 = 6.35$$

If level of education plays a strong role in decisions about family size then some of the observed numbers in Table 6.2 are likely to be very different from the expected numbers (in brackets). On the other hand, if education has no bearing upon family size – if the reality is near to the null hypothesis – the observed numbers should be close to the expected numbers. The **chi-square** (χ^2) technique consists of combining all of the differences between observed and expected observations into a single summary number called the **chi-square** (χ^2) statistic. The **chi-square** statistic is a measure of the distribution of divergence between observed and expected results.

If the observed data are identical to the expected values, χ^2 would equal 0. If the value of χ^2 is larger than would be expected by chance, it is possible to reject the null hypothesis. Critical values of χ^2 are available in mathematical tables (available online) or in software programmes. These convert values of χ^2 into an indication of the probability of a particular χ^2 value occurring purely by chance. (In other words, the probability of the distribution of differences between observed and expected frequencies occurring by chance.)⁵

The general formula for the χ^2 distribution when applied to a contingency table is:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where

r is the number of rows in the contingency table;

c is the number of columns in the contingency table;

O_{ij} are the observed values for each cell (row i column j);

E_{ij} are the expected values for each cell.

The contingency coefficient C is a measure of the probability of obtaining a χ^2 value as large as that found, if it had occurred just by chance, given the null hypothesis of independence.

We can calculate the contingency coefficient, C , from the data in Table 6.2 by means of the χ^2 distribution. From the equation above we have:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

History by Numbers

$$\begin{aligned} &= \frac{(21-24.19)^2}{24.19} + \frac{(23-21.16)^2}{21.16} + \frac{(21-19.65)^2}{19.65} \\ &\quad + \frac{(11-7.81)^2}{7.81} + \frac{(5-6.84)^2}{6.84} + \frac{(5-6.35)^2}{6.35} \\ &= 2.76 \end{aligned}$$

We now find C from the formula:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = 0.18$$

The bigger is C , the closer the relationship but we cannot make comparisons or be precise in predictions because the maximum size of C is affected by the size of the contingency table. In a contingency table with only two rows and two columns the general formula will produce an inflated result for χ^2 and an alternative formula is needed which is also much simpler. To calculate C for a simple 2×2 (2 rows and 2 columns) contingency table with a total of $N \geq 50$ observations, we proceed as follows:

First label the cells as below:

$$\begin{array}{ccc} A & B & A+B \\ C & D & C+D \\ A+C & B+D & N \end{array}$$

Where $N = A+B+C+D$. Then χ^2 will be given by:

$$\chi^2 = \frac{N \left(|AD - BC| - \frac{N}{2} \right)^2}{(A+B)(C+D)(B+D)(A+C)}$$

In this formula, $|AD - BC|$ indicates the ‘absolute value’ of $AD - BC$; that is, we ignore the sign and treat the term as positive even if BC is greater than AD .

C can then be calculated from the formula:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

as before.

Table 6.3 provides a simple example of the use of a 2×2 contingency table in historical research. It is drawn from a comment on land ownership in late nineteenth-century Britain, by Julia Smith. Smith has cross-tabulated the simple binary evidence for

Table 6.3 Businessmen: land ownership cross-tabulation, late nineteenth-century Britain

			LANDED		
			0.00	1.00	Total
BUSMAN	0.00	Count	226	129	355
		Expected count	244.5	110.5	355.0
	1.00	Count	170	50	220
		Expected count	151.5	68.5	220.0
Total		Count	396	179	575
		% of total	68.9	31.1	100.0
<i>Chi-square tests</i>					
		Value	df	Asymptotic significance	
Pearson Chi-square		11.736	1	0.001	

Notes: BUSMAN = 1, businessmen, = 0, non-businessmen

LANDED = 1, have landed wealth, = 0, non-landed wealth

Source: Julia A. Smith, 'Landownership and social change in late nineteenth-century Britain,' *Economic History Review*, 53 (4), (2000), pp. 767–776, p. 770.

businessmen/non-businessmen owning/not owning wealth in land, using data originally collected by Harold Perkin. Her contingency coefficient shows that land ownership and business enterprise were positively related. Note that in the table the chi-square is referred to as Pearson's chi-square after its inventor. The symbol *df* indicates the contingency coefficient, and the asymptotic significance measure is a further indication of the degree to which the result can be relied upon, *statistically speaking*, given the size of the contingency table. The causal connection probably ran from financial success in business to the purchase of land rather than land creating success in business but such arguments are for the historian to make on the basis of the statistical, and other, evidence. They do not exist within the statistical data itself!

The scatter diagram

The **scatter diagram** (also referred to as a *scatter graph* or *scatterplot*) is used with interval or ratio data. It is the simplest visual indication of the possible presence of a relationship between two numerical variables. The formation of a scatter graph involves plotting pairs of the variables against one another on a graph. The pairs of variables can be for cross-sectional or time series data.

Table 6.4 gives time series data for tea and sugar consumption in mid-nineteenth-century Britain, and Figure 6.1 shows the pairs of observations in a scatter graph. If, as in this case, the two variables move roughly together and in the same direction over time

History by Numbers

Table 6.4 Per capita consumption (in pounds weight) of coffee, tea, sugar and tobacco, and an index of average real wages (1850=100), 1850–1865

Year	Wage index	Consumption of:			
		coffee	tea	sugar	tobacco
1850	100	1.13	1.86	25.26	1.00
1851	102	1.19	1.97	26.87	1.02
1852	100	1.27	1.99	29.27	1.04
1853	107	1.34	2.14	30.45	1.07
1854	97	1.35	2.24	33.74	1.10
1855	94	1.29	2.28	30.36	1.09
1856	95	1.25	2.26	28.27	1.16
1857	94	1.22	2.45	29.48	1.16
1858	94	1.24	2.58	34.51	1.20
1859	104	1.20	2.67	34.80	1.21
1860	105	1.23	2.67	34.14	1.22
1861	99	1.21	2.69	35.49	1.20
1862	100	1.18	2.69	35.12	1.21
1863	107	1.11	2.89	35.92	1.13
1864	118	1.06	2.99	36.74	1.28
1865	120	1.02	3.27	36.69	1.30

Source: Based on B. R. Mitchell and P. Deane, *Abstract of British Historical Statistics* (1962), pp. 343, 356.

the points on the graph will roughly line up along a positively sloping line and will be indicative of the presence of a positive relationship.

Similarly, in Figure 3.20, we noted how the relationship between two time series variables, in that case the marriage rate and GDP per capita in the USA, could be represented on a scatter graph demonstrating a positive association and creating food for thought about causation. A further example is provided in Figure 6.2: a scatter graph derived from research on pre-industrial income levels. It gives paired observations between percentage employed in farming, forestry and fishing and average income per person (standardized/converted to 2005\$ for comparison). This is technically cross-sectional data even though there are 182 observations in total for various points in time over 1946–2005, spread over 66 low income countries (average income $\leq \$4,500$) (there is no time series here). The scatter diagram clearly shows an association between agricultural employment/activity and average income levels. This time the association is ‘negative’: the line of best fit between the points slopes downwards from left to right). As the weight of the agricultural sector in an economy gets less, average real incomes tend to rise.

A further example of the use of a scatter diagram, again using transnational data, is found in a recent study of railways, government and export growth in Latin America in the late nineteenth and early twentieth centuries. Figure 6.3 illustrates the positive

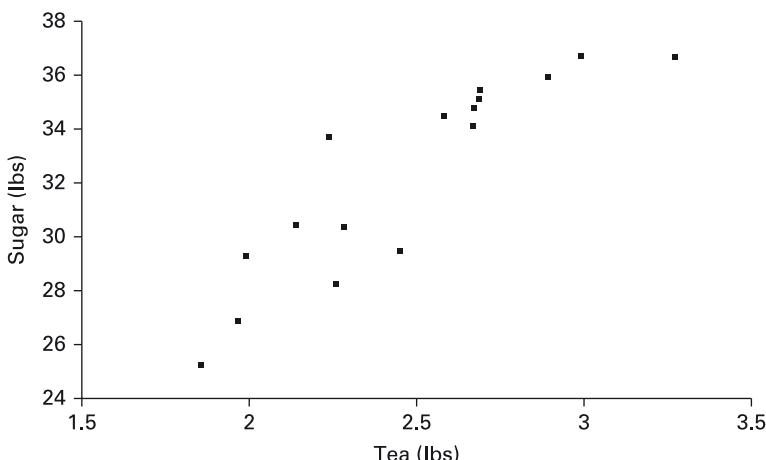


Figure 6.1 Scatter graph showing tea and sugar consumption, 1850–1865.

Source: See Table 6.4.

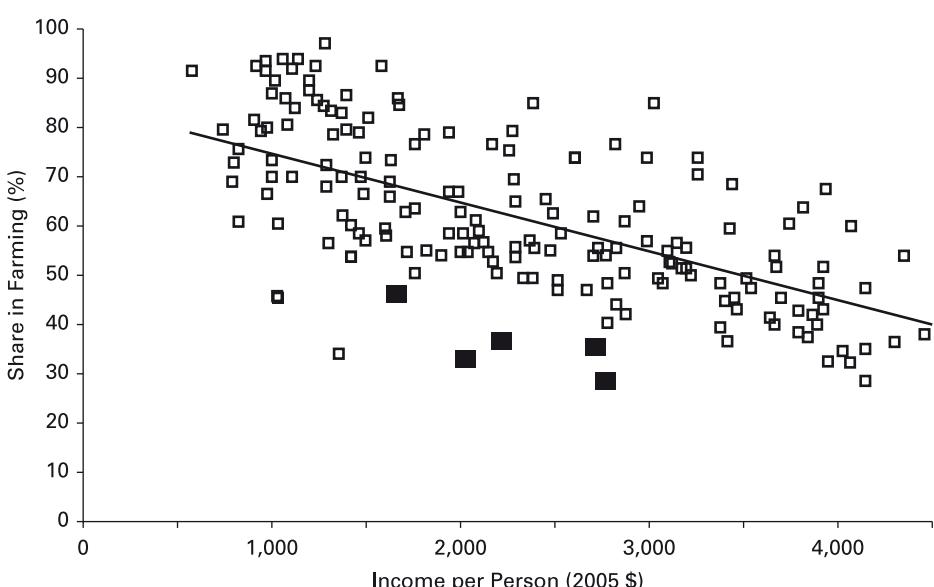


Figure 6.2 Shares in farming, forestry and fishing and real income per person, 1946–2005.

Note: The black squares are the observations from Guyana in 1946, 1960, 1965, 1977 and 1980.

Source: Greg Clark, Joseph Cummins and Brock Smith, 'Malthus, wages, and preindustrial growth', *Journal of Economic History*, 72 (2), (2012), pp. 364–392, p. 365.

association/correlation between railway development (taken as km per capita) and export levels per capita: those countries that experienced higher levels of exports per capita also invested more resources per capita in railways. The Log measure is applied on both scales in order to fit the data range onto the figure in a manageable way. The

History by Numbers

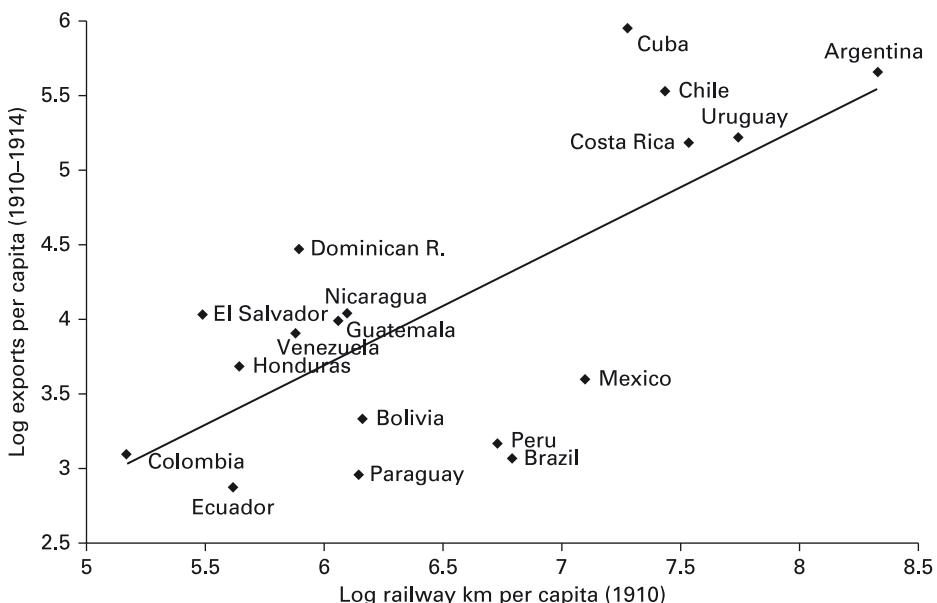


Figure 6.3 Railways per capita and exports per capita in Latin American economies, 1910–1914.

Source: V. Bignon, R. Esteves and A. Harranz-Loncan, 'Big push or big grab? Railways, government activism and export growth in Latin America, 1865–1913', *Economic History Review*, 68 (4), (2015), pp. 1277–1305, p. 1280.

authors go on to use regression analysis (see below) to test whether insufficient government revenues resulted in delayed construction of the railway network in some countries. This article is the subject of an exercise on p. 292.

In drawing a scatter graph or any other sort of graph involving independent and dependent variables (or variables suspected of being dependent and independent); it is the convention to place the independent variable on the horizontal or *x*-axis and the dependent variable or variables on the vertical *y*-axis. Remember: an **independent variable** is one which is suspected of causing the movement of others without in turn being affected by them; a **dependent variable** is one whose variation is seen as being causally dependent on the movement of another variable. Thus in Figure 6.3 railway development was placed in the horizontal axis and exports on the vertical scale.

Figure 6.4 shows different scatter graph shapes with their different indications. Sometimes it is not clear to see at a glance whether there is much of a relationship indicated or not. In this case it is sometimes helpful to draw dotted lines through the median point of each variable. If the majority of points fall in quadrants one and three (formed by the dotted lines) a positive relationship may be indicated. If the majority of points fall in quadrants four and two a negative relationship may be indicated. If the points are widely dispersed across all four quadrants there will be no indication of a

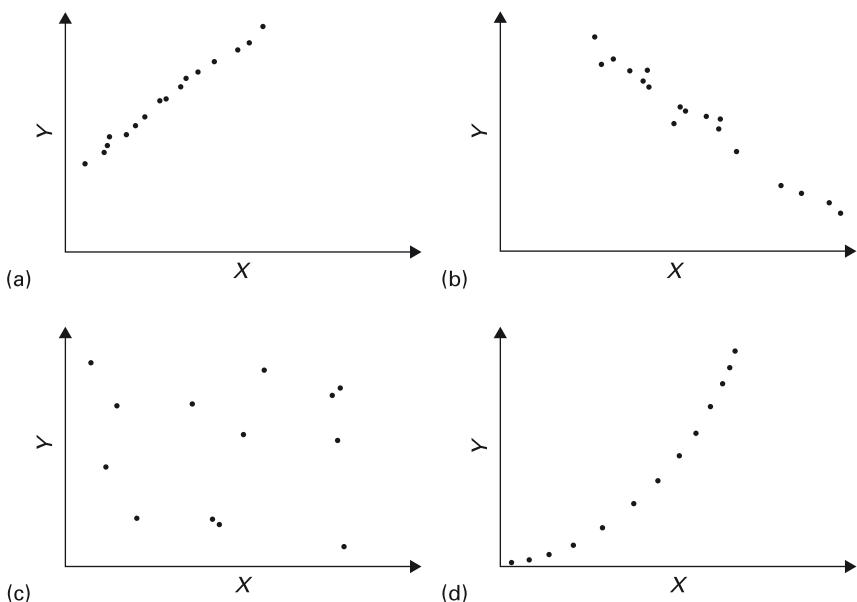


Figure 6.4 Scatter graphs with different indications: (a) positive linear correlation indicated (R value near to +1); (b) negative linear correlation indicated (R value around -0.9); (c) no correlation indicated (R value near to 0); (d) no linear correlation indicated but a curvilinear or monotonic increasing relationship is indicated (R value near to 0 but r_s near to 1).

Note: The correlation coefficient, R , and Spearman's rank correlation coefficient, r_s , are both discussed below.

relationship and this is a signal that it is not worth proceeding with any more sophisticated analysis to measure or test such a relationship.

Scatter graphs sometimes indicate points clustered around a non-linear shape that *may* indicate a non-linear relationship (for example, in Figure 6.4(d)). Spreadsheets and statistical packages can generate scatter graphs with great ease and it is always wise to take this easy first preliminary step in identifying the possible existence of an association between variables before attempting any more complex analysis.

If, after drawing a scatter graph, a linear relationship is suspected it is possible to use statistical techniques to identify it more closely and to measure its strength. The most commonly used tests of statistical association between variables are concerned with linear association and it is these techniques that are the focus of this chapter. It is however important to note that there are many other relationships of dependence which need different techniques. If, for example, each increase of two units in one variable were matched by an increase of 2 squared (4) units in the other, and so on, their plot would be a parabola and not a straight line. Any test for linear association would fail even though the two variables have a very clear association (see the discussion of Spearman's rank correlation coefficient on pp. 180–181).

Dummy variables

Before passing to other methods of identifying and assessing relationships between variables, it is necessary briefly to return to our earlier distinction between numeric and categorical variables. We have seen that the contingency table and the contingency coefficient can handle nominal, ordinal, ratio or interval data. Other methods, and most of the techniques of inferential statistics more generally, can only cope with numeric data. To counter this problem categorical variables can in some cases be re-coded as numerical **dummy variables**.

One of the most common uses of dummy variables arises in coding **dichotomous variables** such as sex. A dichotomous variable is one where there can only be two values. For example, instead of Male or Female one could give a score of 1 to each female, 0 to each male. In a population of 1,000 with 600 men and 400 women the 'mean' of the variable women would be 400/1,000, that is, 0.4. This has an intuitive meaning because 40 per cent of the population is female. In Table 6.3 above, landowning/non-landowning and businessmen/non-businessmen have been coded 1 and 0 respectively, giving numerical character to the qualitative data in order to form the contingency table.

The same strategy can be extended to categorical variables that have more than two values (that is, that are not dichotomous). In studies of voting behaviour, for example, where voters are all Conservative, Labour or Liberal, we would create two new dummy variables, Conservative and Labour, each of which would either take the value 0 or 1. There would be no need to create a dummy variable for the Liberal voters because this information would be captured where cases recorded 0 in *both* the Conservative and Labour columns. If there were four classes of voter, Conservative, Labour, Liberal and Other, one would create three dummy variables leaving 'Others' to be captured by the 0 scores under the other three headings. Similarly, with a larger number of categories such as with the occupations found within an enterprise one could create dummy variables to capture the information numerically.

The advantage of dummy variables is that they can be used in statistical techniques such as correlation and regression (see below) whereas the categorical information upon which they are based cannot.

Correlation coefficient (R)

The correlation coefficient, or R , is also known as Pearson's product moment coefficient of correlation (after its inventor, Karl Pearson: see Chapter 2). The correlation coefficient is the most commonly used measure of association between two variables. It is used to test for the possible presence of a *linear* relationship between two numerical variables. It is wise to plot the data at the outset, in a scatter diagram, to make sure a linear rather than a non-linear relationship is suggested. If a non-linear relationship is present calculating a correlation coefficient will produce an entirely spurious result. If a linear relationship is suspected we adopt the following line of reasoning and action in calculating R :

Step 1

First, we can measure whether a relationship appears to exist between two sets of numerical data by looking at the product of deviations from the mean of each variable for each case.

Step 2

If a relationship exists, the product of deviations will be large and positive, if not they will be near to zero.

Step 3

To get more accurate results, and to facilitate comparisons, we now correct for the type of unit and number of cases, and also for the spread of values by dividing by:

- (i) the number of cases;
- (ii) the product of the standard deviation of each variable.

These steps are incorporated into the formula for R :

$$R = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\left(\frac{\Sigma(X - \bar{X})^2}{N}\right)} \sqrt{\left(\frac{\Sigma(Y - \bar{Y})^2}{N}\right)}}$$

If R is not zero then some sort of relationship may exist.

How strong is the relationship?

The greater C and R are, the closer the relationship. But the drawback of C (the **contingency coefficient**), as we have noted, is that its maximum size (that is, the indication of a strong and predictable relationship) varies with the size of the contingency table. This limits its use.

R is handier because its value always lies between +1 and -1 and its value is not affected by the number of cases or the spread of the distribution. Values of R from different sets of data can therefore be compared:

$R = +1$ if there is a perfect positive correlation

$R = -1$ if there is a perfect negative correlation

The nearer R is to +1 or -1, the closer the relationship (that is, one variable is moving in a fixed way proportionately to the other). In a scatter graph this would show up as points clustering around and along a straight line: with a positive slope if R is near to +1 and with a negative slope if R is near to -1 (see Figure 6.4(a) and 6.4(b) respectively). An R near to 0 indicates no relationship as in Figure 6.4(c).

The form of the relationship

We usually want to know not just that a positive or negative relationship may exist between variables but whether one variable is *responsible* for causing the movement of the other. Statistical tests will never be able to establish this but they can indicate the strength of the association and add weight to a reasoned historical argument about causation. As we have noted, the variable regarded as the dominant or causal variable is usually termed the **independent variable** whilst the other variable is termed the **dependent variable**.

It is sometimes obvious which is which but more usually it is a matter of historical judgement or *hypothesis*. It is important to remember, that judgements and hypotheses about the direction of causality between two variables may be wrong and that the same correlation result would appear whichever direction the causal connection ran. For example we may be interested in examining the hypothesis that wages are related to prices in a particular locality but it would be very difficult to disentangle whether high prices caused high wages or vice versa. Similarly, we may be interested in the extent to which unemployment variations in the inter-war period were determined by the level of welfare benefits in relation to locally available wage levels, that is, by regional variation in the relative generosity of unemployment benefit indicated by the benefits to wage ratio, B/W (where B is the level of benefits and W is locally available wage index). Discovery of a statistically suggested relationship between the movement of the two variables unemployment (U) and B/W could, however, just as easily be a result of the impact of unemployment upon wage or benefit levels.⁶

We should not jump to conclusions about causality on the basis of strong correlation or contingency results without having a good argument, including additional evidence, which will confirm which is the dependent and which is the independent variable beyond reasonable doubt. We must also beware of rushing to conclusions about the causal implications of correlation or contingency results because the movements of two variables may both be dependent upon the movement of a third variable and may not be causally related to one another at all. For example, the rise of both tea and sugar consumption per head of the population in the late nineteenth century may be primarily a result of efficiencies in international trade and shipping or both might result from rising real incomes, rather than one commodity being a major cause of demand for the other. In this context it would be hasty to hypothesize or to closely examine the causal effect of tea consumption upon sugar consumption even though the scatter graph and the correlation coefficient between the two variables may be temptingly positive and strong (see Table 6.4 and Figure 6.1). Sugar was of course used for many other purposes than as an accompaniment to hot beverages, particularly for preserving fruit, and for making jam and desserts.

In an exciting piece of work tracing the growth of books and manuscripts in circulation in Europe as an element in the 'Rise of the West', Buringh and Van Zanden test their hypothesis that there was a strong association between the increase in the numbers of monasteries, as seats of literacy and learning, and the rise of books and reading. They

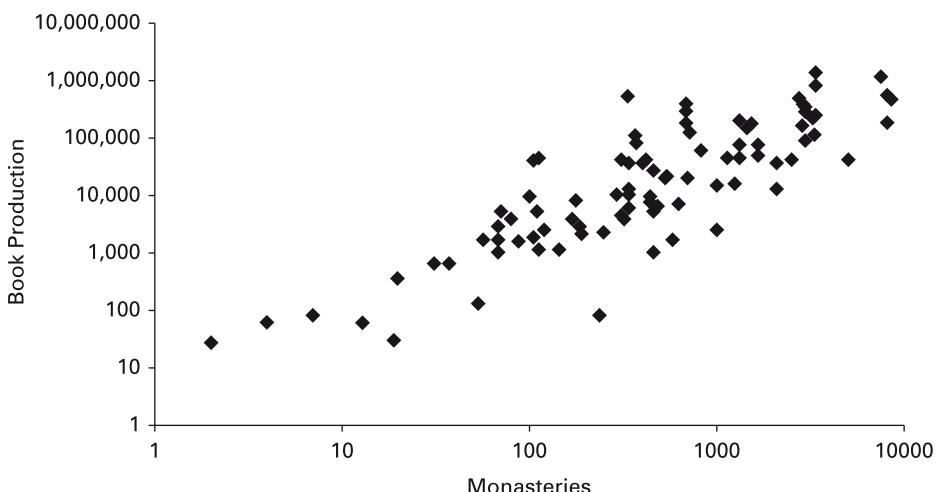


Figure 6.5 Book production and the number of monasteries in Europe, sixth to fifteenth centuries.

Source: Eltjo Buringh and Jan Luiten van Zanden, 'Charting the "Rise of the West": manuscripts and printed books in Europe, a long term perspective from the sixth through the eighteenth century', *Journal of Economic History*, 69 (2), (2009), pp. 409–455, p. 428.

plotted, on a scatter graph (Figure 6.5), estimations of the two variables for a large number of different European places and points in time between the sixth and the fifteenth century. They felt able to conclude that book production was 'to a large extent driven by the number and size of monasteries which was in turn determined by the share of the agricultural surplus that regions and countries directed to this type of activity'.⁷

Lagged results

An important difficulty that occurs with correlations of time series data is that an association may appear not to be present either from time series graphs, scatter graph evidence or from the correlation calculation. However, if we strongly suspect (on the basis of careful historical judgement) that a causal connection may be in operation it may be worth introducing a lag into the series of the independent variable because it is often the case that the causal effect of one factor upon another takes time to work through. The length of the lag that we introduce must have some sound historical reasoning behind it. For example, Wrigley and Schofield's major study of population change in England between 1541 and 1871 suggested that nuptiality was influenced by the movement of real wages. But this apparent causal connection showed up statistically only if a lag of about 30 years was introduced into the real wage series.⁸ As this was such a long lag with little explanation of why real wages may have taken so long to impact

History by Numbers

upon nuptiality a generation or so later, the causal explanation provoked severe criticism. Joel Mokyr, for example, suggested that 'since no-one controls the fertility of his own children let alone his grandchildren ... the lag structure Wrigley and Schofield propose is no explanation at all but a description of the data'.⁹

Another example of results to be obtained from lagging data occurs in research on the determinants of profit rates in the Yorkshire textile industry in the early nineteenth century. As wages were the single biggest input cost in the industry for most of the period, Hudson's hypothesis was that changing wage levels may have had a big impact upon profit rates. No significant correlation showed up between profit rates for various firms in the worsted industry and either money wage or real wage movements when the

Table 6.5 Lagged profit rates in the wool textile industry correlated with wage series for four firms (Clough, Bairstow, Foster and Marriner), 1840–1858

Variables ^a	Period	N	Correlation coefficient	Significance at 5 per cent level
Money wages on:				
Clough profit rate	1845–56	11	0.40	No
Bairstow profit rate	1840–58	18	0.69	Yes
Foster profit rate	1842–58	16	0.47	No
Marriner profit rate	1842–58	16	0.15	No
composite profit rate A	1842–58	16	0.62	Yes
composite profit rate B	1845–56	11	0.67	Yes
Rousseaux real wage index ^b on:				
Clough profit rate	1845–56	11	0.52	No
Bairstow profit rate	1840–58	18	0.76	Yes
Foster profit rate	1842–58	16	0.58	Yes
Marriner profit rate	1842–58	16	0.29	No
composite profit rate A	1842–58	16	0.82	Yes
composite profit rate B	1841–56	11	0.71	Yes
GRS real wage index ^c on:				
Clough profit rate	1844–49	5	0.69	No
Bairstow profit rate	1841–50	10	0.84	Yes
Foster profit rate	1841–49	8	0.51	No
Marriner profit rate	1841–49	8	0.60	No
composite profit rate A	1841–49	8	0.83	Yes

N = Number of observations.

^a Profit rates lagged by one year.

^b Real wage index calculated by means of the Rousseaux price series.

^c Real wage index calculated by means of the Gayer-Rostow-Schwartz price series.

Source: Pat Hudson, *The Genesis of Industrial Capital: A Study of the West Riding Wool Textile Industry, c. 1750–1850* (Cambridge 1986), p. 245.

yearly indices were compared without a lag. But, when the wage series was lagged by a year the correlations were uniformly negative (as one might expect). More surprisingly, lagged profit rates were associated positively with rising real wages but not all of the R values were significant, given the short data runs available.¹⁰ Table 6.5 gives the results of the correlation exercises for different wage indices and different lagged profit rate series. For discussion of the significance of correlation results and the impact of small sample sizes see Chapter 7.

A more recent example of a lagged correlation calculation is provided by a study of the determinants of height in European countries over 1720–1910. The focus of the research is the need to determine how and why the Portuguese became the shortest Europeans. The authors first identify that average height in Portugal was not out of line with that in other European countries in the eighteenth century. But when European anthropometric values began to increase around 1850, the Portuguese seem not to have shared in this rise. In the regressions on height in European countries, height is the dependent variable whilst three sets of lagged data (numeracy, literacy and real wages) and the urbanization rate and poor relief rate (unlagged) are the independent, or proposed explanatory, variables. The lags introduced are decade-long (the result of reasoning about how long the impact of such variables might take to impact upon height). Using such data, literacy and numeracy show a positive impact on height. The authors conclude that delayed human capital formation (defined as the accumulation of education and skills in the population) in Portugal could have had a significant influence on the biological standard of living. The results of the regression analysis are laid out in Table 6.6. The relationship between correlation and regression is explained on pp. 182–184.

Table 6.6 Panel regressions on height in twelve European countries, 1720–1910

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Numeracy	14.38*** (0.008)	6.733** (0.020)	10.71* (0.068)		9.217*** (0.001)	11.70** (0.029)	14.10*** (0.009)
Literacy				6.379*** (0.000)			
Real wages (log)	0.829* (0.087)		1.891*** (0.000)		1.467*** (0.00311)	0.865* (0.0658)	
Urbanization (log)	1.676*** (0.004)	0.970 (0.199)			2.020*** (0.000)	1.980*** (0.000)	
Poor law	2.036** (0.044)	2.577*** (0.002)		2.437*** (0.001)	3.159** (0.012)		
Relative price of protein (log)				0.298 (0.813)			
Infant mortality					0.00867 (0.109)		

(Continued)

History by Numbers

Table 6.6 (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Time fixed effects	Yes						
Country fixed effects	No	Yes	No	No	No	No	No
Weighted	Yes	Yes	Yes	Yes	Yes	Yes	No
Constant	146.9*** (0.000)	159.5*** (0.000)	156.3*** (0.000)	163.2*** (0.000)	159.7*** (0.000)	146.8*** (0.000)	145.9*** (0.000)
Observations	79	165	79	134	157	79	79
Adjusted R ²	0.614	0.745	0.522	0.606	0.583	0.600	0.579

Columns (1)–(3) refer to different numeracy proxies. The last three rows allow for different time periods, population sizes etc.

Source: Y. Stoltz, J. Baten and J. Reis, 'Portuguese living standards, 1720–1980, in European comparison: heights, income and human capital', *Economic History Review*, 66 (2), (2013), pp. 545–578, p. 570.

Spearman's rank correlation coefficient

Several correlation coefficients exist other than Pearson's which is the most commonly used. These are suitable for specific sorts of data, generally where non-linear associations between variables are suspected. One of the most frequently used alternative correlation coefficients in social science research is Spearman's rank coefficient. It was developed for analysing psychological data. Before the advent of computers the Spearman method used to be used to simplify calculations when calculating Pearson's correlation coefficient (for linear association it gives a result quite close to the Pearson coefficient). But Spearman's coefficient can also be used as a measure of the strength of a relationship when variables have a curvilinear association (that is, when they cluster near a line that is curved). Such associations are termed 'monotonic'.

Figure 6.4(d) illustrates a monotonic association where the Spearman coefficient would give a score near to +1 whereas the Pearson coefficient would not and would miss the presence of a reasonably clear association simply because it is not linear. Similarly, Spearman's measure would give a better indication of an association between two variables if the monotonic form was negatively sloped.

The Spearman coefficient, r_s is easy to calculate. Each variable is ranked in size order and then the Spearman correlation coefficient is calculated from the ranks. (Average values are used when the ranking produces tied values). A scatter graph of the relationships between pairs of rankings can be drawn as a preliminary to further analysis. If this shows that the points line up on a concave or convex curve, the Spearman coefficient can usefully be applied.

Table 6.7 shows the sort of original data that would be used in calculating Pearson's coefficient alongside the ranking figures that one would employ to calculate Spearman's.

Table 6.7 Age at death and value of moveable property of clothiers, with corresponding rankings, 1760s

Age at death (years)	Rank	Value of moveable property			Rank
		£	s.	d.	
23	1	27	19	4	3
39	2	31	4	3	7
41	3	25	18	10	2
49	4	25	5	0	1
50	5	31	2	4	6
53	6.5	34	12	9	12
53	6.5	42	0	0	14
54	8	29	1	3	4
56	9	32	10	0	8
57	10	30	5	4	5
58	11.5	33	1	6	9
58	11.5	33	17	6	10
60	13	41	2	8	13
61	14	34	10	7	11

Source: Hypothetical data of the type that might be derived from matching probate inventory values with family reconstitution evidence on age of death.

It details clothiers' ages and their assets in the 1760s. Hypothetical simplified data is used to make the differences between the two methods clearer.

The scatter graph from this exercise is shown in Figure 6.6. This indicates a weak positive correlation (r_s is actually 0.59) compared with an R value of 0.50 (using the original data). This suggests the possibility of a monotonic rather than a linear relationship. The Spearman's coefficient in this case has eased the calculation with little loss of precision in the coefficient.

A recent example of the application of Spearman's rank coefficient from active historical research arises in a study of productivity patterns in Swedish manufacturing industries over 1869–1912. The aim of the study is to analyse whether Swedish industrialization was a 'yeast-like' process whereby various industries expanded simultaneously in response to common stimuli, or whether it was more 'mushroom-like', that is marked by more patchy and random growth, without an identifiable single cause. In Table 6.8 Spearman's rank coefficients are calculated in order to measure stability in the rank order of industrial sectors experiencing productivity growth in various periods. Large coefficients would suggest a sustained pattern in industrial growth. None of the coefficients are particularly large (although two are statistically significant at conventionally acceptable levels indicated by the p-values). The author concludes that the evidence is not sufficiently convincing for a yeast-like development to be identified. It was rather the case that Swedish industrialization had a mushroom-like character.

History by Numbers

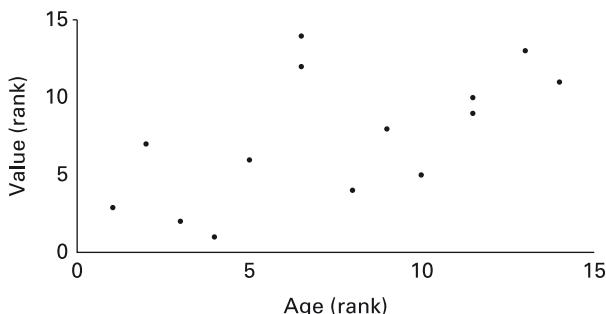


Figure 6.6 Scatter graph of rank of clothiers' ages at death against rank of value of assets.

Source: see Table 6.7.

Table 6.8 Spearman's rank correlation coefficients relating to Swedish industrial growth, 1869–1912

	<i>Spearman's rank coefficient</i>	<i>p-values</i>	<i>No. of industries</i>
1869–79/1879–87	0.4537	0.0199	26
1879–87/1887–93	0.0709	0.7146	29
1887–93/1893–1901	-0.0081	0.9651	32
1893–1901/1901–9	0.3590	0.0269	38

Source: S. Prado, 'Yeast or mushrooms? Productivity patterns across Swedish manufacturing industries, 1869–1912', *Economic History Review*, 67 (2), (2014), pp. 382–408, p. 397.

The regression line

In addition to deciding about the direction of possible causal connection between variables, we may wish to enquire:

1. By how much does X have to alter to produce a change in Y?
2. Can we predict by how much Y would change if the value of X increased by n units?
3. Do changes in variable X explain all changes in variable Y or are other factors involved?

To answer these questions we need to employ further statistical tools starting with the **regression line**. The **regression line** is a line that represents the closeness of movement between two variables. It can be drawn as the line of best fit through the points in a scatter graph.

An example of this is shown in Figure 6.7 which illustrates a hypothesized relationship between the variability of agricultural wages and arable specialization in different English counties. It is part of an argument by Sokoloff and Dollar about the causes of widespread rural manufacturing in England compared with the United States: that this related to the low off-peak opportunity costs for labour in English agriculture.

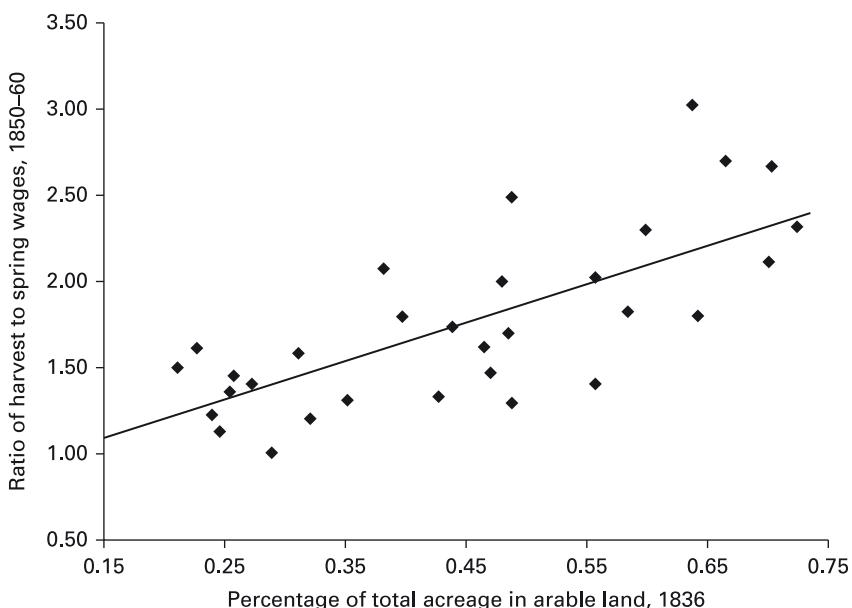


Figure 6.7 Relationship between ratio of harvest to spring wages and percentage of total acreage in arable land, 1836.

Source: K. L. Sokoloff and D. Dollar, 'Agricultural seasonality and the organisation of manufacturing in early industrial economies: the contrast between England and the United States', *Journal of Economic History*, 57, 2 (1997), p. 309.

The **regression line** can be plotted and positioned exactly using the same method that we used to find the position of a trend line of best fit in a time series, that is, by the least squares method and using the same formula (see Chapter 5, pp. 149–151). Remember that the least squares method ensures that the total product of deviations of observations from the line (above and below) is minimized.

The formula for the regression line is:

$$Y = a + bX$$

where

$$a = \frac{\Sigma Y - b\Sigma X}{N}$$

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

and where

- Y is the dependent variable;
- X is the independent variable;

History by Numbers

- a* is the intercept (the point at which the line crosses the *y*-axis);
b is the slope (or gradient).

The regression line represents the best estimate of the relationship between the two variables on the basis of the available evidence. If the fit is good and there are not too many outlying readings, and if we have accurately identified the dependent and independent variables, our predictive capacity, based on the regression line, will be good. (The fit will be best the nearer *R* is to +1 or to -1.) In other words we can use the formula for the regression line and knowledge of its slope to predict by how much *X* has to alter to produce a change in *Y* and to make suggestions about the likely outcome for *Y* of values in the independent variable *X* which lie outside of the available evidence. The equation for the regression line is often called the regression equation.

Since the general form of the regression line is $y = ax+b$ it has become conventional for researchers to talk about '*b*', '*B*' or beta (β) when referring to regressions. '*B*' or '*b*' is also known as the **slope coefficient**. The **slope coefficient** is the number by which the independent variable is multiplied in the regression equation. Sometimes in discussions of regression, the slope coefficient is called a **regression coefficient**.

A firm relationship with *R* nearing +1 or -1, can be **extrapolated** (by using the regression line), beyond the available dataset in order to predict the relationship between the variables in earlier or later time periods (in the case of time series), or across higher or lower ranges of the variables than the information which we have.

An **extrapolation** is an estimation of missing values of a variable or regression line based on the trend apparent in the known values. A regression equation thus becomes a *predictive model*.

A regression equation can also be used to test an explanatory model. If, for example, it is believed that the explanation for the decline of slavery in the plantation areas of the southern United States came with the rising costs of keeping slaves, one could take several sets of plantation or other data indexing costs of upkeep and slave density in different periods or on different plantations and use the regression lines generated to test the model.

The coefficient of determination

We can get a measure of the 'unexplained' element of change in the dependent variable by using the **coefficient of determination**. The **coefficient of determination** represents the degree to which the movement of one variable is associated with variation in another. It is calculated as follows:

$$\text{coefficient of determination} = \frac{(\text{distance between mean and regression estimates})^2}{(\text{distance between mean and data values})^2}$$

In fact this is always equal to R^2 :

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

where

Y is the data value;

\bar{Y} is the mean value;

\hat{Y} is the regression estimate.

R^2 tells us exactly what proportion of the variation in Y we have explained by the regression equation (of Y on X) as being due to the influence of X . Note: if, for example, $R = 0.7$, $R^2 = 0.49$ (that is, only 49 per cent of the variation is attributable to the influence of the relationship). For a dominating influence we must therefore look for values of R greater than 0.7. $R = 0.8$, for example, means that 64 per cent of the movement of Y can be explained by the movement of X .

Examples of correlation and regression analysis in history

To illustrate the use in historical research of the statistical analysis explained above we will now take two examples. The first uses hypothetical data to simplify and clarify the procedures although it is based on research carried out many years ago by a famous American economic historian called W. W. Rostow. It considers the possible relationship between the number of social disturbances and the price of wheat.

Table 6.9 gives information about the annual number of disturbances, the price of wheat, and an index of the business cycle in Britain between 1810 and 1821.

Table 6.9 Economic conditions and social tension in the early nineteenth century

Year	Disturbances (no.)	Price of wheat ^a	Business cycle index
1810	50	105	5
1811	45	95	0
1812	60	125	0
1813	40	105	2
1814	20	75	2
1815	10	65	3
1816	30	75	1
1817	38	95	3
1818	28	85	4
1819	22	75	1
1820	20	65	1
1821	15	55	2

^a Shillings per quart.

Note: In the business cycle index a value of 0 represents a deep depression, 5 a major peak.

Source: Loosely based on W. W. Rostow, 'Trade cycles, harvests and politics, 1790–1950', in *British Economy of the Nineteenth Century* (1948), pp. 123–125.

History by Numbers

The first step in any analysis of data where a linear relationship may be suspected is to hypothesize (on the basis of historical judgement and knowledge) about which of the variables may be dependent and which independent. Looking at the data in Table 6.9 and knowing the context of discussion in Rostow's book the major hypotheses might be as follows:

1. That the number of disturbances may be the dependent variable and be causally related to wheat prices (an independent variable).
2. That the number of disturbances (again the dependent variable) may be influenced by the state of the business cycle (which would reflect unemployment and trade levels), the independent variable.

The next step would be to draw the scatter graphs (Figure 6.8). These will tell us if it is worth proceeding with causal analysis and what sort of relationship appears to be indicated.

We can calculate the correlation coefficients using computer software:

$R = 0.95$ for number of disturbances against wheat prices;

$R = -0.11$ for number of disturbances against the business cycle.

The associated coefficients of determination are, then, $R^2 = 0.90$ and $R^2 = 0.01$, respectively. In other words it appears that wheat prices may account for 90 per cent of the level of disturbances and business conditions may account for 1 per cent of the level of disturbances. The regression equations for the regression lines (of best fit through the scatter graph points) are:

$$Y = -28.7 + 0.71X$$

(where Y = disturbances, X = wheat prices) for the relation between wheat prices and disturbances; and

$$Y = 33.65 - 1.08X$$

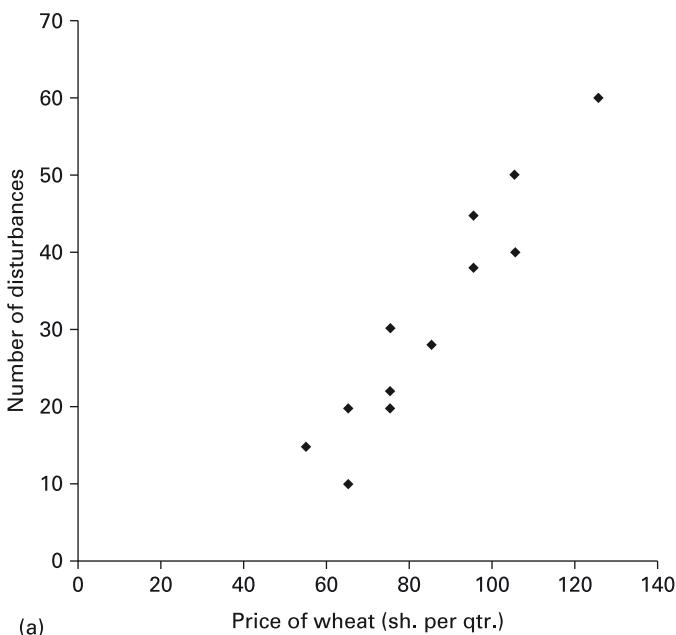
(where Y = number of disturbances, X = business cycle index) for the relation between the business cycle index and disturbances.

The slope coefficients or regression coefficients are thus:

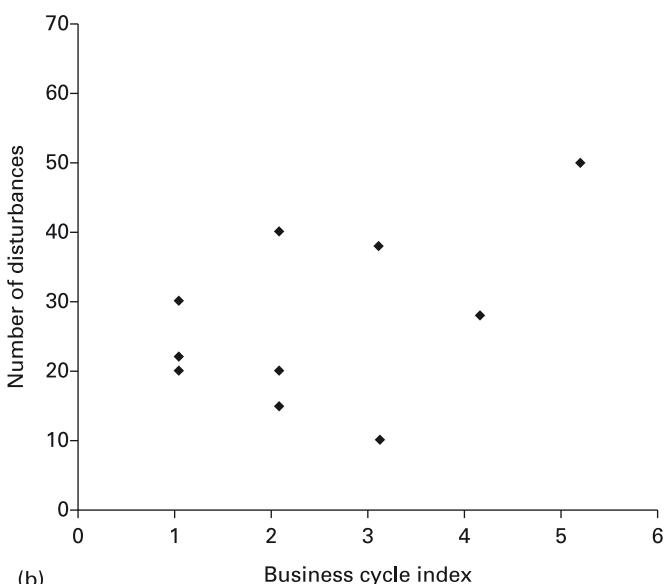
$B = 0.71$ for disturbances against wheat prices

$B = -1.08$ for disturbances against the business cycle index

The significance of the correlation and regression results here is important if we wish to try to generalize from the experience of these few observations. To do this we will need a measure of the probability that the slope coefficients have been thrown up by chance or



(a)



(b)

Figure 6.8 Scatter graph of (a) wheat prices (shillings per quart) and number of disturbances; and (b) business cycle index and number of disturbances, 1810–1821.

Source: see Table 6.9.

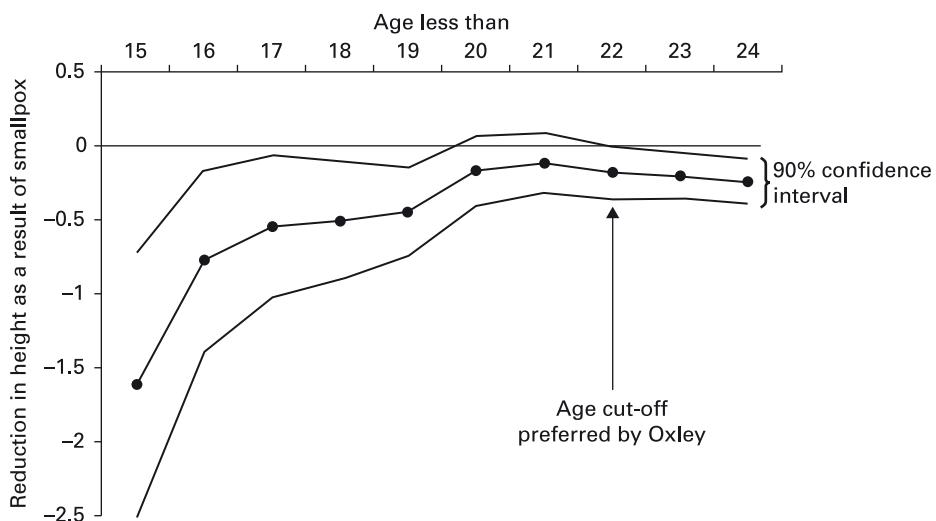
History by Numbers

error arising from the restricted data available. We consider this in Chapter 7. We will also find out (below) that it is possible to run a multiple regression exercise with more than one independent variable, instead of considering the importance of the independent variables separately. The multiple regression would give a more accurate indication of the relative importance of the two influences.

Before leaving the example above, we hope that the data has given some cause for questioning and concern. Even though this is only hypothetical data and has been generated to give a clear example of statistical method, it is important to get in the habit of continuously questioning the reliability, precision and representativeness of figures. How accurate is the measure of the number of disturbances? What sorts of disturbances were included (for example, strikes, riots, demonstrations, rallies)? And to what extent is this really an index of social tension? Do wheat prices stand as a good proxy for living costs, dearth or economic conditions for the majority of the population? If the subsistence or non-market sector was large people may have been insulated from market prices for food by being able to produce their own. Or maybe oats or potatoes were relatively cheap and could be substituted when prices of wheat were high. In addition the business cycle index seems a rather contrived and blunt tool with only five gradations. How has it been created? On what original data was it based? And what problems might it create in an exercise of this kind? There are many pitfalls that might arise if we accept this sort of data at face value.

As our second illustration of linear regression analysis in a historical application we return to a debate we have already encountered (pp. 75–76): that concerning the impact of smallpox on height in nineteenth-century Britain. Tim Leunig and Hans-Joachim Voth carry out a bivariate regression of height against pock marks (Yes/No) using an aggregated sample of several thousand transported convicts, prisoners and others for whom height data is available in London, elsewhere in Britain, and in Ireland. Their results suggest that smallpox significantly reduced stature irrespective of location which conflicts with Oxley's argument that the impact was only significant in London where the population also suffered from other urban health insults and disamenities. Leunig and Voth also suggest that the impact of smallpox on height was greater for younger people, another result disputed by Oxley on the basis of her subsample data. The impact of smallpox on average height by age, according to the 'meta-analysis' of Leunig and Voth, is shown in Figure 6.9. Their results and the relevant *t*-statistics (a measure of statistical significance: see Chapter 7) of the meta regression, compared with the disaggregated results of Oxley are given in Table 6.10.

A further example where regression is this time used to aid time series analysis, is found in Joyce Burnette's study of wages and employment of female day labourers in English agriculture between 1740 and 1860. A scatter graph of female nominal summer wages plotted against time appears to show an upward trend over time but because the regional variations in the data are so great, simply tracing the average wage over time does not give an accurate measure of the time trend of wages. Burnette controls for regional differences in the size and pay of the female labour force by regressing the natural log of the wage on controls for region as well as on time dummies.¹¹

**Figure 6.9** Impact of smallpox on average height, by age.

Source: Timothy Leunig and Hans-Joachim Voth, 'Comment on the seat of death and terror', *Economic History Review*, 59 (3), (2006), pp. 607–616, p. 610.

Table 6.10 Meta-analysis results of pock marks against height regression

	London	English rural	English urban	Irish	Meta-analysis
Oxley's smallpox					
coefficient	-0.440	-0.061	-0.193	-0.143	-0.196
t-statistic	2.94	0.41	1.31	1.34	(2.93)

Notes: We use the Sharp-Sterne implementation of the meta-routine (Hedges and Olkin, *Statistical Methods*). Meta-analysis includes fixed effects, z-value in parentheses.

Source: Timothy Leunig and Hans-Joachim Voth, 'Comment on the seat of death and terror', *Economic History Review*, 59 (3), (2006), pp. 607–616, p. 610.

Multiple regression models

In the real world and in most historical processes a complex web of interacting variables (dependent and independent) is at work. It is often the case that a dependent variable is simultaneously affected by the movement of several independent variables and not just by one. Multiple regression models can be used to investigate associations between the movements of variables where there are several explanatory variables that appear to be operating. Multiple regression enables one to estimate the force of each effect in the presence of other interacting effects. It helps in sorting out which explanatory variables appear to be the most important and which are unimportant.

Statistical (and some spreadsheet) programs enable one to feed in the data for dependent and independent variables and to generate the corresponding regression

History by Numbers

equations and regression coefficients (together with standard error measures which may be necessary to take account of the sampling process: see Chapter 7).

When several variables are taken together in a multiple regression model their regression coefficients indicate the effect of each variable because the other variables in the model are effectively held constant. When each new variable is introduced the coefficient changes. The degree of importance of each independent variable (in impacting upon the dependent variable) can similarly be compared by holding constant the other variables. And the coefficient of determination or R^2 can be assessed for all of the major independent variables combined leaving a measure of the degree of variation that remains unexplained.

Statistical packages make multiple regression manipulations relatively easy but diminishing returns usually set in after just two or three variables have been included in the model. Very unwieldy, not always very helpful, models are sometimes created. Using multiple regression models to make any sort of prediction is a dangerous business. The most important thing that can be achieved with multiple regression models is the ability to sort out which explanatory variables appear to be the most important and which are unimportant.

When encountering research in which multiple regression techniques are used it is a good idea to beware of what can seem to be a sophisticated battery of almost incontrovertible evidence about the importance of an explanatory model.

It is always a good idea to do the following:

1. make sure to think about the quality and reliability of the raw data (as with practically all types of quantitative research).
2. Consider the coefficient of determination or R^2 , and decide whether the model has satisfactorily explained the major causes of change in the dependent variable.
3. Think about whether any potentially important explanatory variables may have been omitted from the model.
4. Ask whether some of the independent variables that have been included may be measuring almost the same thing. If the independent variables in the model are themselves highly correlated, this will distort the analysis.
5. Finally, whether reading historical research that involves multiple regression or undertaking such research for yourself, think hard about the explanatory variables that have been included. Avoid including all the possibilities one can think of, asking the computer to calculate regression equations on every possible combination. As with all statistical work in history the value of the analysis is likely to lie far more in the historical skills applied than in any of the quantitative techniques. The latter are only valuable when applied with thoughtful historical insight, knowledge of the sources, and knowledge of the period and the issues concerned.

A fairly straightforward example of multiple regression analysis is found in a recent study of the characteristics of apprentices in early modern London. Having identified the fact

that apprentices appeared to be significantly younger between 1575 and 1810 than had been the case earlier, the authors examine the possible causes of this age reduction by carrying out a multiple regression analysis. Table 6.11 shows the results of five regression models where the age of binding to an apprenticeship is the dependent variable. The values in parentheses are *t*-statistics (these indicate the degree of reliability that might be attached to the findings, given the sample size, and in relation to the hypothesis being tested). Model 1 covers the full period with the status and occupation of the father and the region of origin as the independent variables. Model 2 introduces a 'company dummy', that relates to the twelve most prestigious London companies whose members monopolized senior municipal offices. The third model examines regions of origin, whether pastoral or arable. The fourth and fifth models distinguish between the seventeenth and the eighteenth centuries. From the results laid out in Table 6.11, the authors conclude that the decline in age of apprenticeship and a greater concentration of age occurred gradually throughout the seventeenth and eighteenth centuries and that this was a trend largely independent of region, background or type of livery company involved. There were however variations associated with region and company. Those from areas more distant from London were apprenticed at older ages whilst those from more prosperous backgrounds were apprenticed when younger. Those whose fathers had died tended to be apprenticed later than others. These findings are important for understanding wider issues about the changing availability of labour in families and households in the early modern period.

Further examples of multiple regressions can be found in Chapter 7 and in the exercises following Chapter 7. Federico Etro and Laura Pagani, for example, examine the market for paintings in Italy during the seventeenth century by considering the impact of canvas size, distance, artist's reputation and other variables on the prices of works of art.¹² A bigger question relating to economic development is posed by Tracy Dennison and Sheilagh Ogilvie in their study of the relationship between the European marriage pattern and economic growth. The European pattern of late marriage, high proportions never marrying and the early rise of nuclear family residence have historically been seen as contributing to the rising economic success of parts of Europe in the early modern period. Dennison and Ogilvie's multivariate analysis of over 4,000 demographic observations covering female marriage ages, lifetime celibacy and household complexity across 39 European countries suggests that the most obvious examples of the European marriage pattern were associated with economic stagnation rather than with economic success!¹³

A final example of the use of multiple regression coefficients is found in the work of G. R. Boyer and Tim Hatton on the determinants of rural–urban migration in the southern counties of England and Wales in the late nineteenth and early twentieth centuries. Their regression models incorporate measures of real and expected income incentives, distance, migrant numbers already in the town, the role of live-in service employments as well as change over time. The results of their main exercise are given in Table 7.4. We leave further examination of their findings until Chapter 7 because we there discuss various ways of measuring the significance of the results.

Table 6.11 Multiple regression analysis of London apprentices explaining their age at entering service

	1600–1799			1600–1699	1700–1799
	1	2	3	4	5
Father yeoman	-.031 (-4.11)	-.011 (-1.33)	-.026*** (-3.52)	-.018 (-1.28)	-.029*** (-3.09)
Father other primary	-.020*** (-2.70)	-.009 (-1.29)	-.016** (-2.12)	-.009 (-0.66)	-.021** (-2.35)
Father manufacturing	-.027*** (-3.91)	-.010 (-1.40)	-.027*** (-3.95)	-.010 (-0.71)	-.034*** (-4.35)
Father distribution	-.030*** (-3.40)	-.010 (-1.11)	-.030*** (-3.39)	-.007 (-0.47)	-.047*** (-4.14)
Father sales	-.048*** (-5.35)	-.025*** (-2.84)	-.049*** (-5.44)	-.024 (-1.43)	-.057*** (-5.56)
Father service	-.054*** (-6.79)	-.035*** (-4.43)	-.059*** (-7.32)	-.048*** (-3.04)	-.055*** (-6.12)
Father professional	-.052*** (-6.23)	-.028*** (-3.36)	-.049*** (-5.96)	-.041*** (-2.72)	-.051*** (-5.14)
Father gentleman	-.032*** (-4.28)	-.012 (-1.61)	-.030*** (-3.99)	-.020 (-1.45)	-.030*** (-3.30)
1625–1649	-.008 (-1.28)	-.010 (-1.59)	-.009 (-1.49)	-.007 (-1.15)	–
1650–1674	-.025*** (-4.08)	-.023*** (-3.70)	-.028*** (-4.45)	-.025*** (-3.88)	–
1675–1699	-.047*** (-7.99)	-.047*** (-7.87)	-.051*** (-8.68)	-.049*** (-7.93)	–
1700–1724	-.048*** (-8.05)	-.049*** (-8.08)	-.054*** (-9.04)	–	–
1725–1749	-.074*** (-11.45)	-.076*** (-11.55)	-.084*** (-14.97)	–	-.025*** (-5.75)
1750–1774	-.106*** (-16.36)	-.103*** (-15.45)	-.118*** (-18.24)	–	-.057*** (-13.08)
1750–1799	-.115*** (-16.25)	-.112*** (-15.44)	-.128*** (-18.10)	–	-.066*** (-12.82)
Southeast	-.007 (-1.40)	-.007* (-1.91)	–	-.011* (-1.71)	-.007 (-1.64)
Southwest	.038*** (7.48)	.035*** (7.10)	–	.031*** (3.82)	.041*** (6.10)
Midlands	.045*** (10.75)	.040*** (9.77)	–	.041*** (5.94)	.043*** (7.65)
East	.001 (0.13)	-.001 (-0.29)	–	-.004 (-0.51)	.002 (0.46)
North	.059*** (8.83)	.057*** (8.71)	–	.069*** (6.91)	.037*** (3.87)
Pastoral			.018*** (4.51)		

Arable			-.022*** (-8.47)		
Father deceased	.013*** (4.87)	.012*** (4.51)	.012*** (4.37)	.023*** (5.47)	.005 (1.35)
Father occupation = Company	.023*** (3.05)	.018** (2.41)	.021*** (2.85)	.030*** (2.75)	.014 (1.37)
Great 12 Company	-.015*** (-5.48)	-	-.016*** (-5.66)	-.023*** (-5.96)	-.004 (-0.93)
Company dummy ^b	N	Y	N	N	N
Constant	2.87*** (320.00)	2.85*** (238.70)	2.90*** (332.62)	2.86*** (189.43)	2.83*** (345.53)
R-square	.07	.11	.06	.05	.06
N	18,214	18,214	18,214	9,435	8,779

^a The regression estimation method is Ordinary Least Squares. The dependant variable is the natural logarithm of age at binding. T-statistics are in parentheses. Sons of fathers without an occupation were excluded from the sample. Apprentices identified as training with a relative are also excluded. 'Pastoral' counties are Devon, Dorset, Gloucestershire, Herefordshire, Somerset and Wiltshire. 'Arable' counties are Bedfordshire, Berkshire, Buckinghamshire, Cambridgeshire, Essex, Hertfordshire, Huntingdonshire, Leicestershire, Norfolk, Oxfordshire, Suffolk and Surrey. Coefficients significant at the 1, 5, and 10 per cent levels are marked ***, ** and *. 'Father labourer' is the omitted father occupation group; 1600–1624 is the omitted quarter-century; and London is the omitted place of origin.

^b The 'Great 12' were the twelve most senior London Companies whose members monopolized senior city offices, such as that of alderman.

Source: P. Wallis, C. Webb and C. Minns, 'Leaving home and entering service: the age of apprenticeship in early modern London', *Continuity and Change*, 25 (3), (2010), pp. 377–404, pp. 393–394.

Non-random error, autocorrelation and multicollinearity

There are various problems to look out for when doing regression analysis. Results can be severely distorted by errors, by misspecifications of relationships and by correlations that may characterize the variables under considerations for spurious or accidental reasons. These complications tend to be exacerbated when time series data are used; they are thus particularly problematic in historical applications. In this section we deal briefly with only three of the complications that can occur. These are probably the most common but unfortunately there are more (that can be followed up in the further reading at the end of this chapter).

Non-random error

All correlation and regression analysis incorporates errors. These errors will not radically affect the substance of the correlation and regression results providing that the errors are random. For this to be true however, the omitted variables must be numerous and each, individually, unimportant. They should also occur in different directions (likely in a large population) so that their combined impact upon the dependent variable is small. The errors should be random and unpredictable, that is they should exhibit no systematic pattern and they should certainly not increase or decrease over time. The trouble is that in many regression models derived from economic theory, especially those involving time series, the error terms do change consistently in their impact. (In technical parlance they are not homoscedastic as in Figure 6.10a but change consistently as in Figure 6.10b.) In production functions for manufacturing concerns, for example (such as the Cobb-Douglas production function), causal factors not included in the model such as organizational efficiency, technological differences between plants and

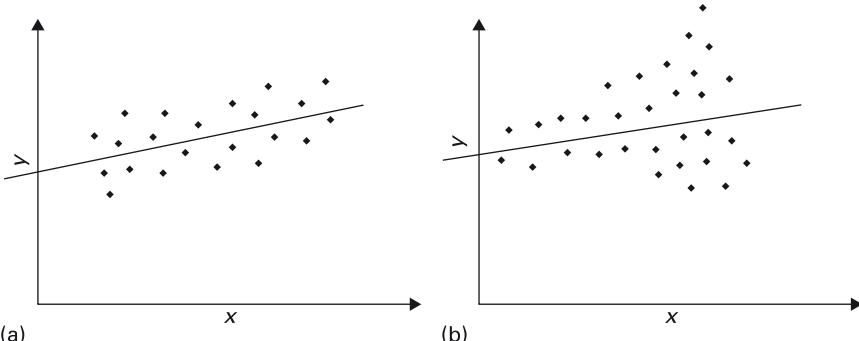


Figure 6.10 (a) Homoscedastic errors; (b) non-homoscedastic errors.

Source: Hypothetical data.

entrepreneurship would vary consistently: it is likely that their influence would be greater in large than in small firms. If the observations in a scatter diagram show increasing variation from the regression line as the values of X and Y increase, estimates for values of Y from values of X will be subject to unacceptable error.

Autocorrelation

A second problem is that the error term may be influenced by its value in an earlier period. This is called **autocorrelation**. It is generally caused by the omission of an important explanatory variable from a regression model or by misspecifying the model (by fitting a linear trend to a curve, for example) or if the effects of random factors such as wars or bad harvests carry over from one period into another. If autocorrelation is present the regression results will not be robust and prediction based upon the results will not be 'efficient'. The errors can be corrected by including the omitted variable, or by re-specifying the model.

Historians frequently make use of a test for autocorrelation called the Durbin–Watson Test, named after its inventors. This involves the calculation of d^* found by the following formula:

$$d^* = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

In this formula e are the residuals (that is, the differences between the regression line estimates and the mean and the data values and the mean for each pair of observations) and t is the time period.

Interpretation of the importance of d^* involves use of a set of conversion tables (available online) but as a rough guide a d^* value of 2 indicates no autocorrelation whilst a d^* nearing either 0 or 4 indicates that there is positive or negative autocorrelation respectively. When using the test both the value of d^* and the interpretation placed upon it should be clearly stated.

Multicollinearity

Multicollinearity can also be a major problem. Multicollinearity occurs when variables are correlated for reasons other than their association with each other. This is frequently the case in correlation and regression analysis of time series data that embody linear or cyclical trends. Correlation will always show up as strong in such cases (R near to ± 1) and regression will *seem* significant. This is because of the time trends in the series and not necessarily as an indication that the two variables may, in any other sense, be associated.

History by Numbers

- If both the variables are rising on trend then R will be near to +1.
- If one variable is declining and the other rising on trend R will near to -1.

If we wish to enquire, for example, about the degree of association between cotton exports and cotton prices in the nineteenth century we would not want our correlation and regression estimates to be influenced by linear trends in the data. The relationship between exports and imports over time is normally very difficult to measure because general growth in the economy (or decline) will affect both and contribute to a common trend.

There is a strong tendency for economic variables to move together over time in prosperity and recession and the increasing use of lagged variables to allow for the delayed impact of a variable has increased the probability of errors caused by multicollinearity. Sometimes the trend elements are clearly visible when the data are graphed independently in time series. But, if not, the Durbin–Watson statistic will again be useful, this time in indicating the disturbing effects of simultaneous trends. If multicollinearity is suspected there are several ways to deal with it.

Dealing with autocorrelation and multicollinearity

If autocorrelation or multicollinearity are identified or suspected there are two common ways of minimizing their impact upon correlation or regression calculations and analysis. The different methods are appropriate for different sorts of distributions and the underlying nature of stochastic, cyclical and trend elements in the time series. These must be identified by careful attention to historical context and complementary historical evidence as well as by analysis of the statistics themselves. Many of the methods are beyond the scope of this volume, although it helps to be aware of them in understanding research that one might read.

The simplest method, which has limited use and is only appropriate for linear data, is to use the **series of first differences** instead of the original time series. Series of first differences are formed by subtracting each value from its predecessor. In Table 6.12 the original series, an index of money wages, has been converted to the series of first differences.

As we saw in Chapter 5, Vincent Barnett uses seasonally adjusted figures of regional prices to consider the efficiency of supply and demand in Soviet commodity markets during the period of the New Economic Policy (NEP). To eliminate autocorrelation and multicollinearity he uses the series of first differences for agricultural product prices. This leaves the figures clear in demonstrating the degree of integration (or otherwise) of markets across the USSR.¹⁴ Table 6.13 is drawn from Barnett's research. It indicates the different results that he obtained from correlations of the original grain price data and the correlations of the series of first differences of grain price data. Quarterly data are used so the sample sizes are small.

Table 6.12 Conversion of a wage index to a series of first differences, 1790–1850

Year	<i>Money wage index (1840 = 100)</i>	<i>Series of first differences</i>
1790	70	
1795	82	-12
1800	95	-13
1805	109	-14
1810	124	-15
1816	117	7
1820	110	7
1824	105	5
1831	101	4
1840	100	1
1845	98	2
1850	100	-2

Source: index derived from P. K. O'Brien and S. L. Engerman, 'Changes in income and its distribution during the industrial revolution,' in R. Floud and D. N. McCloskey (eds), *The Economic History of Britain since 1700* (Cambridge, 1981), vol. 1, p. 169.

In their study of economic warfare in Spain in the early 1940s Caruana and Rockoff use the Durbin–Watson test for autocorrelation in their analysis of the supply of tungsten ore. Tungsten was a vital ingredient in the manufacture of machine tools and armour-piercing shells. Their aim is to consider the impact of US pre-emptive buying of tungsten ore (wolfram) on the open market in an effort to force Germany to pay more for Spanish wolfram and therefore to use less. In estimating the determinants of the supply function for tungsten, especially the price variable, during the period 1941 to 1944, and over the longer term (1911–1955) as a test, they not only use first difference series of (the natural log of) production but also employ Durbin–Watson to demonstrate that the autocorrelation problem is insignificant. Their figures show that the supply elasticity of wolfram was relatively low in the short run, making it a viable candidate for pre-emptive buying (Table 6.14).

In a recent study of the determinants of social policy between 1880 and 1930 Sergio Espuelas demonstrates that inequality did not favour the development of state welfare interventions and social spending between 1880 and 1930 (Table 6.15). He does this by using various independent inequality proxy variables for different countries in least squares regressions against total social transfers through welfare spending as a percentage of GDP.¹⁵ These include income proportion of the top 0.1 per cent, the level of unskilled in the population, differences in the weight of non-family farms in the social structure of countries, levels of elderly in the population, and an index of democracy. Multicollinearity problems with the democracy index caused Espuelas to drop it from the analysis of the impact of inequality on various social spending programmes by type. The explanations that accompany his table 'The determinants of total social transfers as a per cent of GDP', reproduced as Table 6.15, explain how he dealt with the multicollinearity problems. The

History by Numbers

Table 6.13 Aggregate grain price correlations, Soviet Union, April 1924 to October 1927

Grain price correlation, April 1924 to Oct. 1927						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.365	0.906	0.383	0.731	0.759
Caucasus		1	0.523	0.924	0.576	-0.138
Belorussia			1	0.565	0.799	0.574
Ukraine				1	0.565	-0.115
Siberia					1	0.459
First difference grain price correlation, April 1924 to Oct. 1927						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.584	0.792	0.517	0.770	0.410
Caucasus		1	0.693	0.807	0.664	0.187
Belorussia			1	0.695	0.691	0.008
Ukraine				1	0.645	0.067
Siberia					1	0.576
Meat price correlation, April 1924 to Oct. 1927						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.900	0.949	0.911	0.800	0.649
Caucasus		1	0.827	0.867	0.828	0.590
Belorussia			1	0.867	0.724	0.783
Ukraine				1	0.674	0.719
Siberia					1	0.483
First difference meat price correlation, April 1924 to Oct. 1927						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.707	0.860	0.674	0.742	0.133
Caucasus		1	0.610	0.817	0.626	0.331
Belorussia			1	0.412	0.813	0.343
Ukraine				1	0.504	0.214
Siberia					1	0.511
Metal price correlation, April 1924 to Oct. 1927						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.371	0.573	0.660	0.109	-0.004
Caucasus		1	0.113	0.261	-0.168	-0.067
Belorussia			1	0.010	-0.067	-0.225
Ukraine				1	0.193	0.434
Siberia					1	0.091

<i>First difference metal price correlation, April 1924 to Oct. 1927</i>						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.060	0.558	0.598	-0.145	-0.119
Caucasus		1	-0.236	0.074	-0.174	-0.407
Belorussia			1	0.275	-0.010	0.337
Ukraine				1	0.230	-0.026
Siberia					1	0.242

<i>Manufacture price correlation, April 1924 to Oct. 1927</i>						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.459	0.635	0.593	0.585	0.462
Caucasus		1	-0.168	-0.194	0.059	0.136
Belorussia			1	0.818	0.658	0.232
Ukraine				1	0.853	-0.032
Siberia					1	-0.259

<i>First difference manufacture price correlation, April 1924 to Oct. 1927</i>						
	RSFSR (C)	Caucasus	Belorussia	Ukraine	Siberia	Asian Repub.
RSFSR (C)	1	0.666	0.169	0.680	0.618	0.348
Caucasus		1	-0.058	0.274	0.649	-0.040
Belorussia			1	0.547	0.070	0.224
Ukraine				1	0.335	0.423
Siberia					1	-0.356

At 5% level statistical significance reached when $r > 0.532$

(0.572)

Source: V.Barnett, 'Soviet commodity markets during NEP', *Economic History Review*, 48 (2), (1995), pp. 329–352, p. 347.

VIF test referred to here is a measure of the **variance inflation factor**. This quantifies the severity of multicollinearity in an ordinary least squares regression analysis, like this one. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity. The significance tests referred to relate to the statistical significance of the resulting regression measures given the number of observations involved. These tests of significance and their relevance to historical significance are discussed in Chapter 7. One thing to note immediately regarding this study is the great difficulty in getting comparable variables across such vast areas of the globe and across countries having very different recording systems. Given that only 20 observations of the dependent variable are used, caution must also attend the small sample size.

History by Numbers

Table 6.14 Estimates of the supply function of wolfram

Sample	Annual, 1911–1955		Quarterly, 1941:1–1944:4	
Dependent Variable	Natural Logarithm of Wolfram Production	First Difference of the Natural Logarithm of Wolfram Production	Natural Logarithm of Wolfram Production	First Difference of the Natural Logarithm of Wolfram Production
Constant	1.91 (1.59)	0.03 (0.19)	-5.50 (-1.07)	-0.13 (-0.73)
Real price of wolfram	0.71 (3.45)	0.62 (2.82)	1.00 (2.20)	0.64 (1.66)
AR(1)	0.68 (5.94)		0.54 (1.82)	
Adjusted R-squared	0.67	0.14	0.56	0.11
Durbin–Watson	1.74	1.99	1.40	1.38

Notes: Absolute values of *t*-statistics are in parentheses.

Source: Leonard Caruana and Hugh Rockoff, 'A Wolfram in sheep's clothing: economic warfare in Spain, 1940–1944', *Journal of Economic History*, 63 (1), (2003), pp. 100–126, p. 114.

In another example, more thoroughly explored in Chapter 7, Boyer and Hatton's analysis of the determinants of agricultural wages in England, 1866–1912, was, for obvious reasons, accompanied by the d^* (Durbin–Watson) statistic (see Table 7.4).¹⁶

An alternative method of eliminating the impact of autocorrelation upon correlation and regression analysis is to use the *detrended rather than the original series* for the analysis. As was shown in Chapter 5, the trend line for each variable can be calculated and the detrended series is formed by subtracting the trend values from the data values for each variable. A good example of this is given in Table 6.16 which calculates the trend and seasonal components of wheat prices for Winchester College in the early eighteenth century. This leaves the detrended, de-seasonalized series available for analysis. The methods used for de-seasonalizing as well as detrending were discussed in Chapter 5, together with other examples.

If it is thought that seasonal or cyclical components are causing multicollinearity they can also be 'removed' by using various filters calculated for different periodicities. Crafts, Leybourne and Mills used the Kalman filter on their series of British industrial production that exposed an identifiable acceleration in the trend rate of growth during the decades of industrialization. This could not be seen or assessed from the raw figures themselves.¹⁷

Other methods of dealing with autocorrelation and multicollinearity are available depending upon the sort of data involved and their distribution over time.¹⁸ The aim of all is the same: to transform the series into one that is stationary (that is, not distorted by the impact of particular chronological influences).¹⁹

Table 6.15 The determinants of total social transfers as percentage of GDP, 1880–1930

	(1) Least squares	(2) Least squares	(3) Least squares	(4) Least squares	(5) Tobit	(6) Tobit	(7) IV
Top incomes (0.1%)	-0.146*** (0.0455)	-0.151*** (0.0578)					
Top incomes (0.1%)* democracy	0.0284 (0.0302)	0.0589 (0.0381)					
Ratio GDP/unskilled wage			-0.276* (0.155)	-0.399** (0.159)	-0.721*** (0.211)	-0.798*** (0.205)	
Ratio GDP/unskilled wage* democracy			-0.197 (0.195)	-0.166 (0.189)	0.530 (0.320)	0.626** (0.311)	
Non-family farms							-0.0140* (0.00842)
Non-family farms* democracy							-0.00318 (0.00762)
Log(GDP per capita)	0.859*** (0.315)	0.767** (0.334)	0.670*** (0.187)	0.611*** (0.185)	0.776*** (0.217)	0.704*** (0.209)	0.191 (0.209)
Elderly	0.170* (0.1000)	0.0937 (0.112)	0.143* (0.0778)	0.117 (0.0819)	0.179*** (0.0545)	0.168*** (0.0527)	0.0377 (0.0600)
Democracy			0.551 (0.353)	0.582** (0.295)	-0.735 (0.494)	-0.821* (0.477)	0.696 (0.514)
Time trend	-0.0603 (0.0550)		0.00580 (0.0468)		0.0645 (0.0586)		
Constant	-5.855** (2.712)	-4.936 (3.198)	-5.252*** (1.295)	-4.220*** (1.508)	-5.834*** (1.757)	-4.492** (1.913)	-0.353 (1.937)
Time fixed-effects		Yes		Yes		Yes	Yes
R ² /pseudo R ²	0.417	0.424	0.419	0.445	0.276	0.306	0.447

(Continued)

Table 6.15 (Continued)

	(1) Least squares	(2) Least squares	(3) Least squares	(4) Least squares	(5) Tobit	(6) Tobit	(7) IV
Left censored observations					10	10	
Total observations	40	40	75	75	75	75	91

Notes: Regressions 1 to 4 include country random-effects. The democracy variable was dropped in regressions of cols. 1 and 2 because multicollinearity problems were detected, after applying a VIF test. Multicollinearity problems probably arose because of the scant number of observations in these regressions. No multicollinearity problems were detected in the new top-income-shares regressions without the democracy variable or in the regressions in which the ratio of the GDP per capita to the unskilled wage is used as a proxy of inequality. However, in the case of the latter variable, the coefficient associated to inequality in regressions 5 and 6 (Tobit) is higher than in regressions 3 and 4 (least squares). At the same time, the positive effect of democracy disappears in regressions 5 and 6. In fact, in regressions 5 and 6, democracy's effect only becomes positive for high levels of inequality due to the positive correction of the interaction term. This suggests that inequality, democracy, and the interaction term are partially correlated. As I said, this does not involve a problem of multicollinearity (the value of the VIF test is below 10 for all the variables). However, it seems that, in Tobit regressions, inequality is capturing part of the effect that the least-squares regressions attribute to the democracy variable. In the instrumental variable regression, instruments for the share of non-family farms are the lagged values of the *share of non-family farms* and of total social transfers. Robust standard errors in parentheses for regressions in cols. 1 to 6. *** Significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.

Source: Sergio Espuelas, 'The inequality trap. A comparative analysis of social spending between 1880 and 1930', *Economic History Review*, 68 (2), (2015), pp. 683–706, p. 699.

Table 6.16 Separation of trend and cyclical components from wheat prices (in shillings per quart) for Winchester College, 1713–1718

Year and quarter	Wheat price	Trend value ^a	Deviation from trend	Seasonal component ^b	De-trended, de-seasonalized series	De-seasonalized series
1713						
1st	42.67	46.71	-4.04	-0.09	-3.95	42.76
2nd	56.88	45.86	11.02	1.55	9.47	47.41
3rd	49.78	45.01	4.77	0.73	4.04	49.05
4th	46.21	44.16	2.05	-2.19	4.24	48.40
1714						
1st	32.00	43.31	-11.31	-0.09	-11.22	32.09
2nd	32.00	42.46	-10.46	1.55	-12.01	30.45
3rd	32.00	41.61	-9.61	0.73	-10.34	31.27
4th	28.44	40.76	-12.32	-2.19	-10.13	30.63
1715						
1st	46.21	39.91	6.30	-0.09	6.39	46.30
2nd	49.78	39.06	10.72	1.55	9.17	48.23
3rd	42.67	38.21	4.46	0.73	3.73	41.94
4th	35.56	37.36	-1.80	-2.19	0.39	37.75
1716						
1st	39.10	36.51	2.59	-0.09	2.68	39.19
2nd	39.10	35.66	3.44	1.55	1.89	37.55
3rd	40.29	34.81	5.48	0.73	4.75	39.56
4th	33.77	33.96	-0.19	-2.19	2.00	35.96
1717						
1st	43.84	33.11	10.73	-0.09	10.82	43.93
2nd	32.00	32.26	-0.26	1.55	-1.81	30.45
3rd	32.00	31.41	0.59	0.7	-0.14	31.27
4th	32.00	30.56	1.44	-2.19	3.63	34.19
1718						
1st	24.89	29.71	-4.82	-0.09	-4.73	24.98
2nd	23.70	28.86	-5.16	1.55	-6.71	22.15
3rd	26.67	28.01	-1.34	0.73	-2.07	25.94
4th	24.89	27.16	-2.27	-2.19	-0.08	27.08

^a The trend values were calculated from the estimated linear trend equation.

^b The seasonal component was calculated by taking the mean of the deviations from trend for the first quarter of the year, then the mean of deviations for the second quarter, etc. This gave values of -0.08, 1.56, 0.74, -2.17; summing these values gave 0.05, but by definition the seasonal variation should have a neutral or zero effect over the whole year. The seasonal means were therefore adjusted by approximately -(0.05/4) in each case, taking them as -0.09, 1.55, 0.73 and -2.19, which sum to zero, and these were used as the estimate of seasonal variation.

Source: R. Floud, *An Introduction to Quantitative Methods for Historians* (2nd edn, 1979), p. 114; figures derived from W. H. Beveridge, *Prices and Wages in England from the Twelfth to the Nineteenth Century* vol. I (London 1939), p. 82.

Conclusion

This chapter has dealt entirely with **inferential statistics**. These are techniques of statistical analysis that go beyond a description or display of data. Inferential statistics involve making predictions or estimates and testing hypotheses concerning causality. There are many more pitfalls here for the unwary researcher or reader than there are in relation to descriptive statistics but, providing care is taken to be a good historian first, and a ‘number cruncher’ only second, these techniques can be both powerful and useful.

Further reading

- Aron, A. and E. N. Aron, *Statistics for the Behavioural and Social Sciences* (New Jersey 1997), Chapters 11 and 12.
- Clegg, F., *Simple Statistics: A Course Book for the Social Sciences* (Cambridge 1982), pp. 173–186.
- Daly, F., D. J. Hand, M. C. Jones, A. D. Lunn and K. J. McConway, *Elements of Statistics* (Harlow 1995), Chapters 11, 13 and 14.
- Darcy, R. and R. C. Rohrs, *A Guide to Quantitative History* (Westport 1995), Chapters 6, 8 and 9.
- Feinstein, Charles and Mark Thomas, *Making History Count: A Primer of Quantitative Methods for Historians* (Cambridge 2002), Chapters 8–11.
- Foster, Liam, Ian Diamond and Julie Jeffries, *Beginning Statistics. An Introduction for Social Scientists* (2nd edition, London 2015), Chapter 13.
- Hanagan, T., *Mastering Statistics* (3rd edition, London 1997), Chapter 10.
- Haskins, L. and K. Jeffreys, *Understanding Quantitative History* (Cambridge, MA 1991), Chapter 6.
- Marsh, C., *Exploring Data: An Introduction to Data Analysis for Social Scientists* (Cambridge 1988), Chapters 8, 10, 12 and 13.
- Solomon, R. and C. Winch, *Calculating and Computing for Social Science and Arts Students* (Buckingham 1994), Chapters 4 and 6.

CHAPTER 7

SAMPLING AND SIGNIFICANCE TESTING

Sampling is an everyday occurrence and we all do it. You have probably sampled this book by looking at only one or two chapters. Sampling is sometimes done with great care, sometimes rather sloppily without much thought about the selection of the sample or the significance of its character as a guide to wider understanding of the whole. The same is the case with social science and historical research that uses samples. In *Figuring Out Society* Ronald Meek summed up popular attitudes to scientific sampling, as opposed to everyday routine sampling, in an amusing narrative:

There was once a man who did not believe in sampling, and who campaigned against it up and down the country. He emphasized all the dangers . . . pointing out in particular that sampling was necessarily based on probabilities rather than certainties, so that you could never be sure that your conclusions were correct. One day he was due to give a lecture on the evils of sampling in a nearby town. He got up, and went down to breakfast. His egg did not look too good, so he tasted a bit of it, found that it seemed alright and finished the lot. He put his hand outside the door, felt that it was raining, and decided to take an umbrella. He looked in the rack for a magazine to read on the train, thumbed through one or two, found one that looked interesting, and put it in his pocket. When the train pulled into the station he chose the carriage that looked the cleanest and travelled to the nearby town. He went to the lecture hall, and gave his anti-sampling lecture, which was received with rapturous applause by an audience of about a hundred people. ‘How did it go?’ his wife asked him when he got home. ‘Wonderful, wonderful’, the man replied, ‘it’s obvious that there’s a strong feeling in the country against sampling’.¹

Often the historian is presented with a massive **population** of cases and it becomes necessary to take a sample for analysis rather than attempt to look at the whole lot. The word ‘population’ is used here in a statistical, rather than a demographic, sense and means those cases that constitute the full dataset under consideration. The most common reason for using samples, and for applying sampling theory, is to reduce the amount of work to manageable proportions with as little reduction in the accuracy and reliability of research results as possible.

The purpose and procedures of sampling

The sorts of historical research where sampling will be necessary include the following. In each of the examples below the time, cost and/or practicality of studying the full dataset (or population) make this prohibitive:

- Analysis of the language and content of merchants' letters in the long eighteenth century. The sheer volume of surviving business correspondence, and its repetitiveness, suggest the need for sampling. A recent project investigated how merchants with little or no face-to-face contact managed, largely through letter writing, to establish common ground and the basis for trust and reliability in business dealings. For this study a sample of business correspondence from regional archives and in the bankruptcy papers of the National Archives was used. There was no systematic way to sample the vast mass of documents that survive but as far as possible a representative sample of firms of different sizes and types, from different regions and sectors and with different levels of success were included. A range of letters covering different periods within the eighteenth century and different economic and political conditions was also the aim.²
- Study of the households of England and Wales as detailed in the census enumerators' books. In a pioneering study of the late 1970s Professor Michael Anderson of the University of Edinburgh studied households listed in the 1851 census by using a 2 per cent sample.³ He selected the total population in 1 in every 15 enumeration districts (945 in all), which included information about 415,000 individuals. In a more recent study Kevin Schurer and Matthew Woppard took a 5 per cent sample of the 1881 census of Great Britain. In this case they took a random sample (for explanation see below) where each household had an equal chance of being selected as part of the sample.⁴
- Interviews with persons working in car factories in the 1960s: here it would be impossible to trace and interview all cases or even a majority of cases. Huw Beynon's classic study *Working for Ford* was based upon interviews at the Speke factory in Liverpool in the late 1960s. His sample of interviewees was a cluster of contacts amongst shop stewards and rank-and-file workers made largely through the Transport and General Workers Union. The aim of the study was to 'extend beyond its specific context – one assembly plant in Liverpool in the 1960s – to say something about the wider fluency of working lives in general'.⁵
- Study of early modern probate inventories: these are detailed lists of moveable possessions at death and were made in connection with the administration of wills. They are an important source for studying the layout and contents of homes, the balance of industrial and agricultural tools and equipment, debts and credit, and the dissemination of new or more varied consumer goods. Many thousands of inventories have survived across the regions and localities of Britain, in different record offices. They date largely from the mid-seventeenth to the late eighteenth centuries and are sometimes stored in date order, sometimes by name and

sometimes rather haphazardly. Some collections are indexed and some indexes include names and occupations. It would be impossible to look at them all but it is quite feasible to take a sample. Lorna Weatherill's early study of consumption goods and consumer culture took, roughly, a 10 per cent sample from selected repositories to reflect a good geographical spread and to cover differing urban and rural locations.⁶ But in practice it was not always possible to take every tenth document from each box, and some compromise had to be made in selecting a sample (as free from bias as possible) without trespassing on the patience and strength of archivists needing to haul around the boxes. One of the biggest more recent studies of the probate inventories of Cornwall and Kent, by a team of researchers at the University of Exeter, also took a 10 per cent sample, sticking to every tenth record where possible. As the Kent and Cornwall archives have reasonably detailed catalogues it was possible to contemplate choosing a sample which would reflect a good spread of occupations, social classes and gender, but this was rejected in favour of a more systematic or random method which would give all types of inventories a chance of being chosen as part of the sample.⁷

- As a final example we cite a recent study of consumption patterns in Britain during the Industrial Revolution, used as an exercise on p. 227. Sarah Horrell, Jane Humphries and Ken Sneath base their study on a sample of individuals captured in Old Bailey court records because they were victims of housebreaking and burglary. Their study includes detailed discussion of the 'sample' of individuals created in this way. Were the victims of crime representative of wider experience? More importantly the authors are concerned about the extent to which the social status of the crime victims changes over time in the court records. If the later records cover those of higher social status than the earlier records the sample cannot be trusted to convey change in the social ownership of consumer goods (clothing, bedding, watches, and so on) over time. It is often the case that historians have a sample thrust upon them in this way because of the nature of the sources. In these circumstances it is important to justify the extent to which the sample represents wider experience, and continues similarly to represent wider experience over time, if (as here) change over time is a focus of the research.⁸

The purpose of sampling in all of the cases outlined above was to create a viable research project with minimum reduction in the accuracy and reliability with which the research results would represent wider features of the populations concerned. A historian who uses a sample should always make clear that it is just that – a sample. He or she should not argue that the sample results necessarily reflect wider experience or practice in the whole population. The only context in which it is justified to suggest wider implications of the sample results is where it can be demonstrated convincingly that the sample is fully representative of the population as a whole.

Statistical procedures can assist both in sampling (to ensure a representative and unbiased sample is selected) and in analysing the wider significance of sample results.

History by Numbers

Such procedures are designed to allow us to make measurably accurate predictions about the nature of the total population simply from analysis of a much smaller group.

When a historian studies a body of data deliberately obtained by sampling from a larger population, the sampling method should always be discussed and justified in terms of commitment to the goal of representativeness. The size of the sample should also be mentioned and justified because with some sampling methods and in some situations a larger sample may be required to ensure the minimization of bias. The most representative sample possible is an **independent random sample**, and it is this type of sample upon which the foundation of further statistical analysis rests.

Independent random sample

An independent random sample is a sample of cases taken from the total population so that each case has an equal chance of being chosen as part of the sample. The Electronic Random Number Indicator (ERNI) is a good example of machinery introduced to ensure the choice of a random sample, in this case winners in the regular prizes for Premium Bonds (savings certificates issued by the British government since the 1950s). The various ball machines used by the National Lottery are similarly designed to yield a random group of 7 numbers from the 59 available.

Statistical software usually includes a random number indicator, but where no computer is available random number tables can be used to dictate the sample. Random number tables and random number software create a stream of numbers (up to a maximum dictated by the broader population size) which occur entirely at random, that is, each number in the population has an equal chance of being chosen as part of the sample. Both the random number indicator and random number tables select cases at random from the number available until the required sample size is reached. Only by selecting a sample in this random way can analysis of the sample be heavily relied upon accurately to predict the nature of the population as a whole, and for this to be so the random sample must contain at least 100 cases.

Experience has shown, and statistical theory confirms, that a well-chosen sample (that is, a sample chosen to avoid bias, but especially one which is independent and random) can reveal a full range of experience and can generate predictably accurate measures for the population as a whole providing there are 100 or more cases in the sample. The historian's urge to collect evidence from as many cases as possible may thus be misplaced. The 101st observation gives much less additional precision than the 100th. Statistical theory shows that the proportion does not matter to the accuracy of the sample, what matters most is the selection of an unbiased sample of 100. To be precise the accuracy of a random sample varies inversely with the square root of the absolute number in the sample. We can say this because of our knowledge of the properties of the normal distribution, to which we turn on pp. 212–214 below.

Obviously, any method of selecting or sampling cases that can be justified as being representative will provide a basis for some sort of estimate of the nature of the total

population but only independent random sampling gives us a means of judging *exactly* how close our sample results are to those which would be obtained from the population as a whole.

Sampling theory and method are based on the assumption that an independent random sample is obtainable, but, in reality, a true random sample is seldom used in historical research. What tends to happen is that a random sample is impossible to obtain, so another sort of sample is used *as if it were random*. This must involve some discussion of whether the researcher believes the sample to be similar to a random one. In other words, biases need to be discussed. If the researcher goes on to use statistical procedures appropriate only for a truly random sample, it must be made explicit that a departure has been made from strict statistical practice. This must be justified by explaining the nature of the sample and by explaining how one can be confident that it is representative of the population and near to random.

Where a population is large and the historian decides to take a sample, it is often not practical or not possible to take a true independent random sample. In fact, the only circumstance where a true independent random sample is usually taken from a very large dataset is when the dataset as a whole has been fed into a computer and is machine readable (as with the 1881 census sample mentioned above). That way random number software can be applied to extract the random sample. With large datasets, scattered in separate boxes in archives or on many reels of microfilm such as with merchants' letter books or probate inventories, it would be entirely impractical to extract random cases from the thousands available.

Systematic and stratified samples

Most often, instead of a random sample a systematic sample is taken of a proportion of the population. Alternatively, a stratified sample is selected. A systematic, proportional sample is formed where the historian takes a certain proportion of cases to represent the whole and where these are selected systematically as a way of avoiding bias. Usually, every 10th or 100th case is taken, but this depends on the size of the population, the required sample size and the nature of the research. Where such a sample is selected the historian must be careful that the method of selection has not predisposed the sample to bias, especially where, for practical reasons (such as illegible or missing documents), it is not possible to stick strictly to every 10th or 100th case. Bias may also occur simply from sticking to a fixed periodicity within the sample. If records exist in date order, for example, the researcher must be careful that any sort of cyclical or seasonal variation in the cases does not coincide with the periodicity of selection of the sample.

To assist in avoiding bias a system of regularity is often introduced from an initially random choice. For example, if a 2 per cent sample is required from 100,000 names, the first is selected at random from the first 50 names (50 because the population is 50 times greater than the sample size; that is, $50 \times 2,000 = 100,000$). If the 23rd name is picked the names are then selected at regular intervals until the sample is complete (for example,

History by Numbers

23rd, 73rd, 133rd, and so on). The initially random starting point is what is generally regarded as characterizing a systematic sample, though the term is used more loosely to denote any proportional sample with a regular periodicity that has been justified in terms of its probable lack of bias.

A stratified sample is one that is deliberately selected so that the various groups or strata in the population are represented – as nearly as possible in proportion to their distribution in the population as a whole. The assumption behind stratified sampling is that the strata are quite different with respect to the variable being studied. This often occurs with a mixed urban and rural sample or with a population that has clear divisions along status or class lines. Stratified sampling may appear attractive to the researcher at first sight but is rarely easy to achieve as one needs to know reasonably accurately the distribution of the whole population between strata. If this knowledge is subject to error then so too will be the sample. In addition, the imposition of fixed criteria for the selection of cases can be problematic if the population changes character over time or if the strata have greater fluidity of classification, meaning or character (with respect to the variable being studied) than the historian has assumed. In his oral history of the Edwardian period Paul Thompson tried to include respondents from different occupations and social strata. As no further statistical analysis was hanging on the accuracy of the sample, it was not vital whether the range of respondents reflected the overall distribution of the population.⁹ Sometimes, a sample from just one stratum is used because the researcher wishes to study the strata in the population that is judged to exhibit more variation. This can occur with, for example, urban population in relation to a variable such as housing type. As long as the rationale is explained and the results are not then used to infer anything about the population as a whole this is perfectly acceptable.

The characteristic of random, systematic and stratified sampling is that every individual has a known probability of being included in the sample, providing the rules governing the selection of such samples are followed to the letter.

Other sorts of samples

Other samples frequently used in historical research come about not through the selection of cases from a large dataset but because the historian has to work with the sample that is most easily available to him or her.

A **cluster sample** is formed when a sample is selected on the basis of the ease of access to a particular group within the population. Street-corner interviews are usually given as the most obvious example of cluster sampling in social science research. These can be justified in terms of their representative nature providing it can be argued that the variable being studied has no relationship to the likelihood of someone being on a street corner at a particular time of day. Cluster samples are often preferred by researchers when they believe that most of the variation in the factor being studied takes place within rather than between clusters. In this context, the cluster can be justified by both ease of access and representativeness. In oral history, interviews are sometimes conducted

in one or two old peoples' homes, or amongst a network of friends or neighbours. Documentary sources may also provide a cluster sample where papers survive only in, or for, one location or social group or for a limited number of locations or social groups. Surveys very commonly are forced to use cluster samples of various kinds because, even if the net is cast wide, those who agree to take part in an interview or written survey are self-selecting and may be a particular sort of person with a particular sort of experience that is different from the mass of the population who have no interest in participating in research. Sometimes the researcher takes the self-selected cluster sample and re-weights it to reflect the stratification of the population as a whole, but this has the danger of introducing additional biases rather than eliminating the problems and should generally be avoided in favour of a full exploration of the biases inherent in the cluster.

A cluster sample may be regarded as one of a number of types of samples of convenience. In historical work, such samples are very common. Other samples of convenience, used by historians, include surviving samples and those that are better documented than others (as with the ownership of consumer goods by crime victims, detailed above).

A **surviving sample** is used where the bulk of records have not survived but where the historian wishes to say something about the nature of the population as a whole on the basis of information from the surviving cases only. For example, with the use of business records, diaries, autobiographies, letters, wills, and many other documents, the researcher is at the mercy of what has survived. With such research it is necessary carefully to explore the ways in which the surviving sample may be biased. Autobiographies might survive for the more literate and successful individuals in society or because they were written for publication (and therefore got printed and preserved) rather than for private use. They are likely on balance to give a rosier picture of experiences through the life course than for those whose stories have not survived, as a recent optimistic study of living standards, political and social conditions during the Industrial Revolution has demonstrated.¹⁰ In the case of business records the larger more successful firms are often better represented in documentary sources than are smaller more ephemeral businesses whose stories end in bankruptcy rather than expansion. The wave of business histories of the 1950s and 1960s gave heroic accounts of large and successful firms which for a time may have biased our understanding of British business people and their efficiency and of the commercial climate within which they functioned. Later research based on bankruptcy cases, which provided a more representative sample, produced results that give a more accurate picture of the performance and operation of businesses as a whole.¹¹ Another example of surviving sample bias arises from the biographies of inventors that were collected by the Victorian compilers of the original *Dictionary of National Biography*. Later researchers have relied heavily upon such sources in studying the social backgrounds and the activities of innovators but a recent study has highlighted the 'pitfalls of prosopography' by demonstrating the partiality and the omissions that have only been partly rectified by additions to later editions of the work. It is also suggested that studies of American inventors that rest on similar sources run the risk of perpetuating bias in the same way.¹²

History by Numbers

A *better-documented sample* is often used where the records relating to a sample of cases contain the information required for the research (or fuller details) but where the majority do not. Use of a better-documented sample can result in a richer history being written but it is necessary to contemplate the distortion that may have been introduced by using such a sample if one's concern is to describe and analyse the population as a whole. Better-documented samples often give undue prominence to notorious or exceptional cases. It is often tempting when researching court cases, for example, to spend much longer on those with fuller evidence in depositions. This can yield interesting histories of such cases but these must be qualified by suggesting how well-documented cases may have differed from those, for the same or different crimes, where the records are shorter and less informative. A recent study of small family businesses in the north-west of England between 1760 and 1820, carried out at the University of Manchester, used court records and wills to examine wider characteristics. Over 400 wills enabled the researchers to run a limited number of statistical queries that generated answers with a reasonable level of reliability concerning the characteristics of the population of small firms as a whole. The much smaller sample of surviving court records gave a detailed picture of individual concerns, especially regarding family decision-making and the disposal of family fortunes, but their non-random occurrence and survival gives rise to some concerns about the ability to generalize beyond the sample itself.¹³

Sampling error

All samples introduce elements of bias and distortion that should be discussed by the researcher as openly as discussion of the pitfalls and biases of the sources themselves. We must be careful to think about and allow for these, particularly if our concern is to use the sample in order to say something about the population as a whole. An independent random sample of over 100 cases reduces distortions to a minimum, but even an independent random sample is subject to random **sampling error**: error in representativeness caused by chance or the 'luck of the draw'. Other sorts of samples have their own peculiarities and difficulties. Most are not suitable for further statistical analysis and the results of research upon such samples must be evaluated solely upon the evidence that they reveal about the sample itself. In these circumstances it is possible only to generalize, impressionistically, about the population as a whole. Other samples, especially independent random samples and certain proportional samples (which can be argued to be little removed from random), allow further statistical analysis in order to calculate the certainty with which we can rely on sample results to give us an accurate picture of the character of the population as a whole.

The normal distribution

Further statistical analysis of samples involves the use of **probability theory** (sometimes termed 'error theory') and the properties of the normal distribution (sometimes known

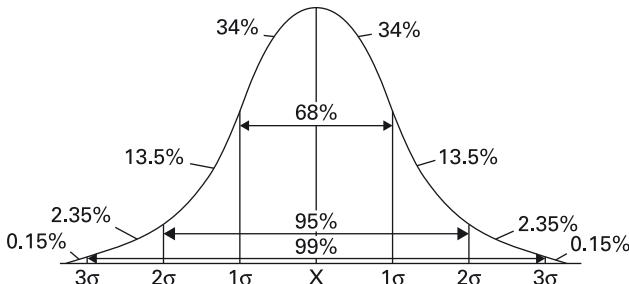


Figure 7.1 The normal distribution. Note: σ = standard deviation; X = mean, median and mode.

as the ‘error distribution’). The analysis revolves around estimating the sampling error arising from an independent random sample and deciding if it is tolerable for one’s purposes. A sampling error is the difference between the ‘true’ value of a characteristic within a population and the value estimated from a sample of that population. ‘Error’ occurs because *no* sample can be expected exactly to represent the population from which it was drawn.

The **normal distribution** is a particular kind of frequency distribution: an ideal type. We encountered the normal distribution in Chapter 4, p. 112. It has two properties:

- The mean, the median and the mode are all the same;
- A constant proportion of cases lie between the mean and multipliers of the standard deviation from the mean:
 - 68.2 per cent fall between one standard deviation above and below,
 - 95.46 per cent fall between two standard deviations above and below,
 - 99 per cent fall between three standard deviations above and below.

The normal distribution is represented in Figure 7.1.

The distribution of sample means

Where does the concept of the normal distribution fit into sampling theory and technique? Let us assume that we repeatedly take random samples of N cases from a population where:

- the mean of the population is μ (pronounced mu),
- the standard deviation is σ (pronounced sigma).

Provided that the number in the sample, N , is greater than or equal to 100 ($N \geq 100$):

- the frequency distribution formed by the sample means will be a normal distribution (known as the sampling distribution). It will always be normal, no matter what sort of frequency distribution was formed by the population itself.

History by Numbers

We can then say:

- the mean of the sample distribution will be the same as the population mean (that is, μ);
- the standard deviation of the sample distribution (the ‘standard error’) will always be $\frac{\sigma}{\sqrt{N}}$

Bearing in mind the properties of the normal distribution:

- there is a 68.26 per cent chance that the mean of a random sample will be in the range $\mu \pm \frac{\sigma}{\sqrt{N}}$
- there is a 95.46 per cent chance that the mean of a random sample will be in the range $\mu \pm \frac{2\sigma}{\sqrt{N}}$

Estimation of the population mean and standard deviation from a sample

We can use the facts outlined in the previous section to estimate the population mean, μ , and standard deviation, σ , from the random sample. This is possible as it can be shown that a good estimate for $\frac{\sigma}{\sqrt{N}}$ is:

$$\frac{s}{\sqrt{N-1}}$$

(Where s is the standard deviation of the sample.) Thus in 95.46 per cent of all samples the population mean will lie in the range:

$$\bar{X} \pm 2 \left[\frac{s}{\sqrt{(N-1)}} \right]$$

Where \bar{X} is the sample mean.

Samples and populations: some examples

A seemingly straightforward example of the use of sampling to provide estimates of population characteristics is given by Floud.¹⁴ His example concerns the use of a random sample of 100 female marriage ages drawn from parish registers, to estimate the mean

age of marriage in a town. A random sample of 100 women is taken from the list (that in these days would certainly be electronically stored and amenable to random number software for selecting the sample). The mean age of first marriage, \bar{X} , is found to be 27 years, with standard deviation 2.2 years.

Applying what we know about sampling theory, probability and the properties of the normal distribution, we know that in 95.46 samples out of 100 the sample mean will lie in the range $\mu \pm 2\sigma/\sqrt{N}$, which is equivalent to saying that in 95.46 samples out of 100 the population mean will be somewhere in the range $\bar{X} \pm (2\sigma/\sqrt{N})$. Since we do not know σ we use s/\sqrt{N} – 1 as an estimate of σ/\sqrt{N} . We therefore know that, for this example, in 95.46 samples out of 100 the population mean will be in the following range:

$$\begin{aligned}\mu &= \bar{X} \pm 2 \frac{s}{\sqrt{(N-1)}} \text{ years} \\ &= 27 \pm 2 \frac{2.2}{\sqrt{100-1}} \text{ years} \\ &= 27 \pm 0.44 \text{ years}\end{aligned}$$

The population mean will therefore be between 27.44 and 26.56. The range around the mean, ± 0.44 years is known as the 95 per cent confidence interval because we can have approximately 95 per cent confidence that the population mean will lie in that range.

From the statistical point of view there is little wrong with this example. From a historian's point of view it does, however, raise some useful questions such as how were the marriage ages of the women in the sample derived from the parish registers? Also, over what period of time is the study located and how was the random sample selected? Most parish registers for periods pre-1800 give no direct evidence of age of bride. This can be calculated only from linking marriage to baptism entries, being certain of the correct match and then allowing an estimated time for the birth–baptism interval. The baptism evidence for all brides is unlikely to be found, so that it would almost certainly be impossible to get a true independent sample of female marriage ages. Floud's example might have been less questionable had he suggested that the female marriage ages were randomly drawn from Civil Registration evidence, which is available in Britain from 1837 and which includes the date of birth of the bride. There is, however, no indication in Floud's example of the period of the study, which would need to be finite to allow a random sample of entries to be collected.

Another sort of example of the use of sample results to infer wider characteristics in a population might be the selection of over 100 households from the census enumerators books (CEBs) for a large city for 1881. Statistics from the sample such as average size of household, average age of head of household, average age of household resident, proportion in household not born in the city, and so on could all be considered in relation to the confidence intervals that we might accept for the population as a whole. Of course, in practice, even using random number tables or random number software where CEB transcripts are available in machine-readable form,¹⁵ a researcher might need to diverge

History by Numbers

from strict random selection. This might occur if the random cases included non-households such as hotels, lodging houses, hospitals, prisoners or other institutions. If the research concerns households these selections would need to be rejected in favour of the next random number of the household nearest to the random number of the institution; in either case a departure from strict random selection would have been made and would need to be explained and justified in relation to the maintenance of a representative example.

Difference-of-means test

The difference-of-means test is used where we want to compare two samples, sample 1 and sample 2, taken from the same population at different time periods. Our two samples may have different means and standard deviations, but we will want to make sure that this difference really does reflect change in the population as a whole rather than just the result of the sampling process. Again, we make use of the properties of the normal distribution. The difference-of-means test makes use of the fact that if we take a large number, N , of samples, where $N \geq 100$, from two populations and calculate the difference between the means of each pair of samples, the sample distribution of the differences will itself be a normal distribution.

The sample distribution will have a mean equal to the difference between the two population means. Its standard deviation or standard error will be:

$$\sqrt{\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)}$$

where subscripts 1 and 2 indicate the sample in question. From the normal distribution we know that there is a 95.46 per cent chance that the difference between the means of the two samples will be in the range:

$$\pm 2\sqrt{\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)}$$

of the difference between the means of the population. We thus know that only in 4.54 per cent (100–95.46) of samples would we be likely to have a difference of means greater than two population standard deviations from zero.

These facts allow us to calculate a measure, z , of the extent to which the differences between the sample means exist only by chance; z is the difference of means divided by the number of pooled standard errors. Since the population standard deviation is unknown the sample standard deviation are used as proxies.

The formula for z is thus:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1} \right)}}$$

In Floud's example of marriage ages he suggests that a mean age of 27 years and a standard deviation of 2.2 years might be compared with measures a century later of 26.5 years and 1.6 years for the same variables, for a repeated sample of 100 cases. On the surface, such a comparison might suggest that the average age of marriage in the wider population had declined by six months over the century. But such a 'shift' is small enough to have been the result purely of the errors produced by sampling. The first sample may have overestimated the average age of marriage and the second may have underestimated it. In Floud's example we can consider this by calculating z :

$$z = \frac{27 - 26.5}{\sqrt{\frac{2.2^2}{99} + \frac{1.6^2}{99}}} \text{ years} = 1.83 \text{ years}$$

Since z is 1.83 years we know that there is a greater than 4.54 per cent chance that the difference between the sample means (six months) is a result of using the samples. Using tables or software that reflects the shape of the normal distribution we can be exact in saying that there is a 6.73 per cent chance that the difference between the samples was the result of sampling and not the result of the two populations being any different. Whether we think this a too large a risk is a question of historical as well as statistical judgement, however. Statistical convention suggests that anything greater than a 5 per cent chance should not be readily dismissed, but the researcher is in charge and must make the decisions on the basis of these figures and not just relate them as a statistical fact.

The difference-of-means test generates the percentage probability that the different sample results were thrown up purely by the sampling process. If the result is small (for example, 5 per cent or less) one can have some confidence that the population as a whole has changed over the period, but it is up to the researcher to decide whether a 5 per cent chance is small or large, tolerable or intolerable, depending upon his or her purposes.

The significance of sample results

The decision on whether to accept that sample results are an accurate estimate of population characteristics depends on how certain one wants to be. It is a historical not a statistical decision or judgement. If the central thesis of a research project rests heavily on such an estimate one would need to be very confident, accept only a small sampling error and seek further complementary and supporting evidence. If

History by Numbers

only a subsidiary part of the research rests on this evidence one might accept a larger sampling error.

Statistical significance merely tells one the probability that the sample results will be an accurate reflection of the population as a whole. The chance taken in accepting sample results is usually expressed as statistical significance at the 10 per cent, 5 per cent, or 1 per cent levels. For example, if a result is said to be significant at the 5 per cent level it simply means that there is a 5 per cent or less chance that it was produced by the sampling process.

The measure of significance is done by the so-called ***t-test*** which, like much in probability theory, also takes advantage of the properties of the normal distribution, this time in estimating what proportion of sample results will lie at the outer limits of the error distribution. Only 5 per cent of the sampling distribution will lie more than two standard deviations away from the population distribution itself.

The *t-test* is a hypothesis-testing procedure in which the population variance is unknown; it compares *t* scores from a sample to a comparison distribution called a *t* distribution to give levels of statistical significance for sample results. It is useful for dealing with small samples and was pioneered by William Sealy Gosset (1876–1937) for testing the quality of small samples of beer at Guinness, where he was employed.¹⁶

A recent example of historical research employing sampling, discussing and testing for bias in samples and applying tests of the statistical significance to determine the level of bias is provided by Horrell, Humphries and Sneath, whose article was briefly mentioned at the beginning of this chapter. They are concerned in their study to ensure that their sample of victims of housebreaking and burglary in the late seventeenth and early eighteenth centuries (drawn from Old Bailey court proceedings) does not change in its overall occupational representation over time. They are especially keen to test whether they are not oversampling those of higher status in the earlier decades (a possibility because of the way in which private prosecutions were brought). This is important because they are trying to demonstrate the changing social spread of ownership of different sorts of consumer goods over time. They do not want their results to be contaminated by a biased sample. To check for this they conduct a number of statistical tests (shown in Table 7.1) which demonstrate to their satisfaction that, although the mean status of victims varies a little over time, there is no clear time trend evidence. Note the use of correlation coefficients, the chi-squared test and regression for this purpose: to compare the spread of occupations across the decadal samples. The *t-test* is then applied to the results. Whether we accept that these results, and their statistical significance, are sufficient proof of lack of bias is of course a matter for the historian as reader.¹⁷

The significance of correlation and regression results

It is possible to test the significance of correlation and regression results in a similar way in order to allow for the probability that results derive only by chance. With small

Table 7.1 Occupational status over time

	1750–1	1760–1	1770–1	1780–1	1790–1	1800–1	1810–11	1820–1
No. of thefts	51	15	123	75	94	127	155	140
No. where occupational status known	23	7	45	31	52	73	101	69
% in each category:								
Titled	4	–	11	3	–	3	2	3
Professional	4	57	13	10	6	1	3	16
Paperwork	4	–	4	–	4	7	3	4
Non-food seller	39	14	24	10	23	21	35	10
Jeweller	–	14	7	7	4	8	9	9
Innkeeper	–	–	13	13	14	11	12	10
Food seller	9	14	4	7	4	10	10	10
Clothing maker	13	–	9	3	14	15	9	9
Trades	17	–	4	19	8	7	3	10
Army/servant	9	–	2	16	12	1	6	3
Casual	–	–	7	13	14	16	9	16
Mean status	50.2	70.0	58.1	38.9	42.9	45.1	50.9	46.7
% female occupations	4.3	14.3	13.3	9.7	11.5	6.8	5.9	8.6

Notes:

Correlation coefficients:

Pearson's R –0.063 (0.21 significance)

Spearman's rank –0.046 (0.36 significance)

Chi-squared:

	1750–1781	1790–1821
Jeweller and above	50.9%	43.7%
Innkeeper and below (sample)	49.1% (106)	56.3% (265)
χ^2	1.636 (0.20 significance)	
<i>Regression:</i>		
Status rank =	52.658 (14.15)*	–0.782 time trend (–1.25)

Adjusted $R^2 = 0.001$, $F = 1.567$, $n = 401$

t-ratio in parentheses.

* Significant at 10% level or higher.

Source: Sarah Horrell, Jane Humphries and Ken Sneath, 'Consumption conundrums unravelled', *Economic History Review*, 68 (3), (2015), p. 840.

History by Numbers

data runs a higher correlation coefficient is required for statistical significance than if the data runs are long. Table 7.2 shows the results of running correlations between profit rates in the West Yorkshire wool textile industry during the industrial revolution and export levels. Table 7.3 gives the correlation results for profit rates against wool prices. In this research the author had suspected that export levels and wool prices may have been important determinants of profitability. Because of the availability of the evidence some of the data runs are very short, and although some high correlation figures have been produced this is not always sufficient, given the short data run, to produce a statistically significant result. A more senior or discerning scholar would probably not have bothered to calculate correlations on such short and ‘blunt’ datasets. Fortunately, Hudson’s entire thesis and academic career did not depend on these results!

From the data in Tables 7.2 and 7.3 it was possible to suggest that profitability in the woollen industry may have been significantly influenced by export variation and by raw wool prices. The data runs were too short and the data itself insufficiently robust to engage in regression or multiple regression analysis.

Tables 7.4 and 7.5 give further examples of the ways in which historical work incorporates measures of the statistical significance of correlation and modelling results. Table 7.4 indicates the determinants of changes in agricultural wages in England and Wales, 1866–1912, under three possible sets of conditions, which are modelled in columns 1, 2, and 3.¹⁸ The *t*-statistics are given in parentheses, which is a common way of displaying them. Some confusion might arise from the use of the letter *t* in the specification of the variables in the left-hand column, but here *t* stands for time. One time period is indicated by *t*, and *t*–1 indicates that the series has been lagged by one time period (in this case one year).

On the basis of the results in Table 7.4 Boyer and Hatton argued that economic conditions in urban areas had a strong influence on short-run wage changes in agriculture. In the same article they also considered the determinants of rural–urban migration in southern counties, and Table 7.5 gives results for five variables, over three different decades, as incorporated into models to produce the figures in numerical columns (1)–(3). From this they were able to suggest that economic incentives, especially the expected income gap between countryside and towns, were important in migration, but an even more important factor was prior migration (measured as migration stock in the models), which reflected human networks of communication and assistance. These are very interesting results but the pitfalls of carrying out such exercises given the available data are numerous and the degree of error, which may be introduced by the modelling process is great. The pitfalls are thoroughly discussed by the authors themselves, which is good professional practice.

Table 7.2 Correlation of profit rates with export levels in the West Yorkshire wool textile industry for seven firms (Hague, Cook & Wormald; Illingworth; T. & M. Bairstow; Foster; Marriner; Broadbent; Clough), 1822–1858

Variables	Period	N	Correlation coefficient	Significant at 5 per cent level?
Blanket exports from UK, by volume, on:				
Hague, Cook & Wormald profit rate	1822–55	34	0.55	Yes
Hague, Cook & Wormald profit rate	1840–55	16	0.83	Yes
Illingworth profit rate	1828–33	6	0.87	Yes
Export of wool goods from Great Britain, at current prices, on:				
Hague, Cook & Wormald profit rate	1822–29	8	0.55	No
T. & M. Bairstow profit rate	1825–29	5	0.86	No
Export of wool goods from United Kingdom, at current prices, on:				
Hague, Cook & Wormald profit rate	1826–58	33	0.44	Yes
Illingworth profit rate	1828–33	6	0.68	No
Yarn exports, by volume, on:				
Foster profit rate	1842–50	9	0.76	Yes
Marriner profit rate	1842–50	9	0.70	Yes
Total cloth exports, by volume, on:				
Broadbent profit rate	1840–50	11	0.60	Yes
Foster profit rate	1842–50	9	0.67	Yes
Marriner profit rate	1842–50	9	0.70	Yes
Clough profit rate	1845–56	12	0.51	No

N = Number of observations.

Source: Pat Hudson, *The Genesis of Industrial Capital* (Cambridge 1986), p. 242.

History by Numbers

Table 7.3 Correlation of profitability and wool prices in the West Yorkshire wool textile industry for the company T. & M. Bairstow, 1840–1858

Variables	Period	N	Correlation coefficient	Significant at 5 per cent level?
Price of Lincoln half-hogs, on:				
T. & M. Bairstow profit rate	1840–58	19	-0.61	Yes
Composite profit rate	1842–58	17	-0.56	Yes
T. & M. Bairstow profit rate	1845–50	11	-0.66	Yes
Composite profit rate	1842–50	9	-0.67	Yes

N = Number of observations.

Composite profit rate = average rate in worsted branch based on three other firms.

Source: See Pat Hudson, *The Genesis of Industrial Capital* (Cambridge 1986), p. 243.

Table 7.4 The determinants of change in the agricultural wage in England and Wales, 1866–1912

	(1)	(2)	(3)
Constant	-0.075 (3.07)	-0.084 (2.99)	-0.058 (3.71)
Change in price (t)	0.166 (4.59)	0.159 (4.22)	0.167 (4.61)
Change in price ($t-1$)	0.097 (2.47)	0.091 (2.26)	0.092 (2.42)
Urban/rural wage ($t-1$)	0.291 (3.53)	0.308 (3.62)	
Urban employment rate ($t-1$)	0.214 (2.79)	0.215 (2.80)	
Urban or rural wage × employment rate ($t-1$)			0.242 (3.95)
Union dummy, 1872–6	0.017 (2.89)	0.019 (2.85)	0.015 (2.83)
Time		0.0001 (0.54)	
ρ	0.509 (3.26)	0.542 (3.55)	0.456 (2.87)
R^2	0.745	0.747	0.742
RSS	0.0028	0.0028	0.0028
Durbin–Watson statistic	1.859	1.904	1.871

Note: t -statistics are given in parentheses; in the left-hand column, t = time; $t-1$ indicates the series has been lagged by one time period (one year).

Source: G. R. Boyer and T. J. Hatton, 'Migration and labour market integration in late nineteenth-century England and Wales', *Economic History Review*, 50, 4 (1997), p. 722.

Table 7.5 The determinants of male migration rates from southern counties in Great Britain to six urban destinations, 1870s–1890s

	Dependent variable: log migration rate		
	(1)	(2)	(3)
Constant	2.19 (4.68)	2.25 (4.82)	0.71 (0.40)
Distance	-1.02 (7.37)	-1.01 (7.38)	-1.04 (7.41)
Wage gap	0.51 (3.88)		0.46 (3.29)
Expected income gap		0.48 (4.38)	
Migrant stock	0.49 (7.63)	0.49 (7.72)	0.47 (7.32)
Service employment			0.51 (0.85)
D^{1870s}	0.13 (1.75)	0.08 (0.95)	0.13 (1.64)
D^{1880s}	-0.40 (4.87)	-0.44 (5.19)	-0.51 (3.40)
D^{1890s}	-0.62 (6.56)	-0.64 (6.84)	-0.74 (4.42)
Origin-county dummies	yes	yes	yes
Destination-county dummies	yes	yes	yes
R ² statistic	0.880	0.881	0.880
N	452	452	452

Note: *t*-statistics are given in parentheses; the dependent variable and all explanatory variables except the dummy variables are defined in logarithms; *N* = number of observations.

D^{1870s} = decade of 1870s, etc.

Source: G. R. Boyer and T. J. Hatton, 'Migration and labour market integration in late nineteenth-century England and Wales', *Economic History Review*, 50, 4 (1997), p. 712.

Table 7.6 is from Botticini's article on Tuscan dowries, which is included as an exercise at the end of this chapter. Note the mix of dummy and interval variables in the model and the asterisks that flag up those *t*-statistics which are statistically significant at the different levels. The table includes two different specifications of the model. Column 1 places more weight upon the difference between the bride's and groom's ages and has a more significant regression result for the importance of the bride's age in determining dowry size. The second specification allows for the groom's age, which gives a much less significant coefficient for the impact of the bride's age on dowry size. As with most exercises of this type, much historical acumen and energy must go into the specification of the model and in discussing the advantages and disadvantages of using different combinations of causal factors. Even more important is the need for wise discretion in discussing the real significance of a variable that has shown up as statistically significant. The *F* statistic, referred to in Table 7.6, is the ratio of the between-group estimate of the population variance to the within-group estimate of the population variance. In other words, it is a further indication of the comparison between sample and population variance. Botticini uses these results to argue that dowry values were positively correlated with the age of brides and that the parents of girls who married down the social scale gave bigger dowries than those who married up.¹⁹

History by Numbers

Table 7.6 Estimates of the dowry function (dowry is the dependent variable), Cortona, 1415–1436²¹

	1		2	
	coefficient	t statistic	coefficient	t statistic
CONSTANT	18.51	0.32	5.87	0.09
YEAR dummy	15.46	1.46	15.67	1.48
GROOM REMARRIED dummy	-14.33	-0.71	-15.14	-0.73
BRIDE REMARRIED dummy	86.75	3.28***	86.52	3.26***
GROOM'S AGE			4.01	1.3
GROOM'S AGE squared			-0.03	-0.87
AGE OF THE GROOM'S FATHER	-0.68	-1.25	-0.69	-1.27
BRIDE'S AGE	8.59	2.33**	5.57	1.39
BRIDE'S AGE squared	-0.19	-2.45**	-0.16	-1.93*
AGE OF THE BRIDE'S FATHER	-0.09	-0.2	-0.05	-0.12
AGE DIFFERENCE	3.29	2.02**		
AGE DIFFERENCE squared	-0.06	-1.46		
GROOM'S HOUSEHOLD'S WEALTH	0.01	4.72***	0.01	4.71***
BRIDE'S HOUSEHOLD'S WEALTH	0.01	6.14***	0.02	6.19***
AGRAGRUP dummy	-0.53	-0.04	-0.14	-0.01
AGRUPDOWN dummy	33.3	1.51	33.1	1.5
AGRUPUP dummy	98.25	5.26***	95.7	5.14***
NONAGRNONAGRDOWN dummy	109.43	6.12***	109.59	6.1***
NONAGRNONAGRUP dummy	75.89	3.26***	76.25	3.25***
NONAGRDOWNDOWN dummy	98.3	4.9***	98.51	4.89***
NONAGRDOWNUP dummy	63.21	3.69***	63.78	3.71***
NUMBER OF CHILDREN IN GROOM'S HOUSEHOLD	-4.5	-1.27	-4.35	-1.23
PERCENTAGE OF DAUGHTERS IN GROOM'S HOUSEHOLD	45.32	1.4	44.48	1.37
NUMBER OF CHILDREN IN BRIDE'S HOUSEHOLD	-8.28	-2.91***	-8.43	-2.95***
PERCENTAGE OF DAUGHTERS IN BRIDE'S HOUSEHOLD	-11.05	-0.59	-11.72	-0.62
RESIDENCE dummy	26.36	2.19**	25.55	2.11**
R ² statistic	0.68		0.68	
Adjusted R ² statistic	0.65		0.65	
F statistic	19.18		18.99	
Number of observations, N			224	

* Significant at the 10 per cent level.

** Significant at the 5 per cent level.

*** Significant at the 1 per cent level.

Source: M. Botticini, 'A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town, 1415–1436', *Journal of Economic History*, 59, 1 (1999), p. 115.

A recent example of the application of the *t*-test and *T*-statistics in a piece of social history research is provided by Patrick Wallis, Cliff Webb and Chris Minns' study of the age of apprenticeship in early modern London, discussed in Chapter 6. A further recent illustration is provided in Bishnupriya Gupta's study of son-preference and marriage in India in the twentieth century.²⁰ Gupta uses data from the 1931 census to study the extent to which rates of marriage were influenced by regional shortages of females occurring as a result of son-preference. She finds a positive correlation between the sex ratio at age 0–5 in the population and the proportion of men never marrying. This is also the case when the aggregated sex ratio is used. Northern India had the highest sex ratio and the highest proportions of never married men in the population. To gauge the significance of her findings in relation to varied regional cultures Gupta uses the south as her reference point and gives each region a dummy variable taking the value of 1. The statistical significance of the impact of the sex ratio on marriage rates in the North is clearly demonstrated in the resulting *T*-statistics (see Table 7.7) but the historical significance is one that the writer must establish with the cogency of her overall argument.

Table 7.7 The sex ratio and the marriage rate of men, India 1931: *T*-statistics of the region effect

Explanatory variable	Dependent variable			
	Sex ratio in 0–5 age group	Sex ratio in 15–20 age group	Proportion of single males in 45–50 age group	Proportion of single males in 45–50 age group
Sex ratio 0–5				
West	0.02 (0.86)	0.08 (1.71)	0.00 (0.6)	0.00 (0.43)
East	−0.00 (0.02)	−0.03 (0.65)	−0.01 (1.33)	−0.01 (1.32)
North	0.04 (2.16)	0.18 (4.64)*	0.03 (5.08)*	0.03 (4.25)*
Central	−0.01 (0.39)	0.04 (0.92)	0.00 (0.10)	0.00 (0.03)
N = 31				

Notes: The reference region is the south. Each region is a dummy variable taking the value 1. *T*-statistics are shown in parentheses.

* Significant at the 99% level.

Source: Bishnupriya Gupta, 'Where have all the brides gone? Son preference and marriage in India over the twentieth century?', *Economic History Review*, 67 (1), (2014), pp. 1–24, p. 16.

Conclusion

Sampling and the analysis of samples are frequently used in the social sciences. In fact, much of the modern apparatus of sampling theory and significance testing, building upon the work of R. A. Fisher, was developed in sociological research. For the historian, sampling theory and significance tests can be very useful where there are large datasets and where it is possible to take a random or near random sample. Most often, however,

History by Numbers

samples are forced upon historians because of the non-random survival of evidence or by the practical exigencies of collecting a dataset. In all circumstances the best practice to adopt is to examine the sample closely and openly to discuss the degree to which the sample is representative of the population as a whole. Finally, and most importantly, one must always consider the reliability of the sample results not only in relation to tests of statistical significance but also in relation to the nature of the research project, its aims and the centrality of the sample to the overall results.

Further reading

- Aron, A. and E. N. Aron, *Statistics for the Behavioural and Social Sciences* (New Jersey, CN 1997), especially Chapters 4–9.
- Daly, F., D. J. Hand, M. C. Jones, A. D. Lunn and K. J. McConway, *Elements of Statistics* (Harlow 1995), Chapters 2–5, 7, 8 and 10.
- Darcy, R. and R. C. Rohrs, *A Guide to Quantitative History* (Westport, CN 1995), Chapters 2, 4 and 5.
- Feinstein, Charles and Mark Thomas, *Making History Count: A Primer of Quantitative Methods for Historians* (Cambridge 2002), pp. 117–224.
- Hannagan, T., *Mastering Statistics* (3rd edition, London 1997), Chapters 3, 4 and 8.
- Haskins, L. and K. Jeffrey, *Understanding Quantitative History* (Cambridge 1988).
- Solomon, R. and C. Winch, *Calculating and Computing for Social Science and Arts Students* (Buckingham 1994), Chapter 5.
- Ziliak, Stephen T. and Deirdre McCloskey, *The Cult of Statistical Significance: How The Standard Error Costs Us Jobs, Justice and Lives* (Ann Arbor 2008); also at www.press.umich.edu/pdf/9780472070077-fm.pdf (accessed 30 September 2015).

Exercises for Chapters 5, 6 and 7

Horrell, Sarah, Jane Humphries and Ken Sneath, 'Consumption conundrums unravelled', *Economic History Review*, 68 (3), (2015), pp. 830–857.

1. What are the purposes of this study?
2. Why should one be wary of taking Old Bailey proceedings as a reliable record of crime?
3. Why might one be wary of the content of the sentence preceding note 36?
4. Why is care taken to examine the nature of the samples of crimes selected?
5. How convincing are the authors' arguments about the suitability of the Old Bailey records for their purposes?
6. What is an 'a priori decision' (p. 837), why is it important to avoid this and how is it avoided in this case?
7. What are the pitfalls of taking the second-hand value placed on a stolen item as an indication of the price?
8. What is CAMSIS and why is it used here?
9. What statistical tests are undertaken in Table 3 and what is their purpose here?
10. Table 4 has some interesting indications of changes in fashion and availability of goods but why might one be wary of the data relating to 1680–1741 in particular (and in terms of comparison with the later figures)?
11. How and why do the regression results correct for changes in price and real wages over time?
12. Explain the findings in Figure 2 in relation either to watches or to stockings.
13. What do the regression results relating to 'domestic comfort items' demonstrate?
14. Why have feather beds been chosen as an object of study and why was it necessary to include information from every theft of feather beds between 1750 and 1820? Does this unsettle you with respect to the conclusions here?
15. To what extent has the article addressed the consumption conundrum outlined in the first sentence?

Odell, Kerry A. and Marc D. Weidenmier, 'Real shock, monetary aftershock: the 1906 San Francisco earthquake and the panic of 1907', *Journal of Economic History*, 64 (4), (2004), pp. 1002–1027.

1. Through what monetary mechanism did the San Francisco earthquake of 1906 contribute to the international financial instability that led to the panic of 1907?

History by Numbers

2. How did the gold standard supposedly work to stabilize international financial flows?
3. What other explanations (apart from the earthquake) might account for the unusually high British gold exports to the USA in late 1906?
4. What is the purpose of the range and standard deviation measures used in Table 2?
5. What statistical exercise is used to rule out the claim that British speculation in the US stock boom created the gold outflows from Britain to the US? Is this convincing?
6. What data do the authors use to document the insurance claim payments and why were many of these delayed until the autumn of 1906 and beyond?
7. What are the periodicities on the horizontal scale of Figures 1 and 3 and why have these been chosen?
8. Why did so much gold flow out of and into New York in 1906?
9. Write a short narrative of the events depicted in either Figure 4 or Figure 5.
10. How and why was the New York panic spread to Europe?
11. Most of the statistical analysis in this article takes the form of graphical representation of comparative time series of gold flows, industrial production and interest/discount rates. How convinced are you by the analysis of these graphs and are there any further statistical exercises that you think might have been attempted?

Barnett, Vincent, 'Soviet commodity markets during NEP', *Economic History Review*, 48 (2), (1995), pp. 329–352.

1. How can correlation analysis of prices assist in indicating the presence or absence of integrated markets?
2. What are the main problems in carrying out a straightforward test relating the correlation coefficients of regional prices to levels of integration?
3. Why is it preferable to use price series collected by the Conjunction Institute, rather than Gosplan Prices?
4. What is a 'free market'? And why would a high measure of integration fail to prove that markets were free?
5. How does the author decide what correlation coefficient justifies the conclusion that a market is unified?
6. Why is distance of the market from the source of goods likely to be a determining factor in price convergence in many cases? What sorts of goods would most likely be affected in this way? How do the findings from bazaar prices in particular confirm the impact of distance (p. 341)?

7. Why does Barnett suggest that Ozerov's method of measurement of integration for grain markets is faulty?
8. What explanation does Barnett give for his finding that foodstuffs markets appear to have been more integrated than industrial goods such as metals?
9. Discuss the application of first difference series to the data in Table A2.
10. Why is the high inflation of sugar prices likely to affect results for the markets in that commodity?

Gorsky, Martin, 'The growth and distribution of English Friendly Societies in the early nineteenth century', *Economic History Review*, 51 (3), (1998), pp. 489–511.

1. What factors may have led to growing friendly society numbers and memberships in the Industrial Revolution period?
2. How does the author set out to examine these factors?
3. Why are the author's hypotheses (p. 490) not really hypotheses?
4. Discuss the strengths and the weaknesses of the Poor Law surveys of 1803 and 1813–1815 in relation to the author's purposes.
5. List five possible determinants of the regional variation in levels of friendly society membership shown in Tables 2 and 3.
6. Might the statistics in Table 4 have been more clearly represented using a different visual technique?
7. What factors might account for the uneven distribution of female friendly society membership geographically (Table 4)?
8. Assess the evidence for the author's argument that friendly society membership depended more upon the level of urbanization than the level of industrialization.
9. Discuss the use of correlation evidence in arguing that friendly society membership was also related to the *pace* of urban growth. (Appendix 1, columns D–F.)
10. What is the aim of carrying out a correlation exercise to consider the relationship between friendly society membership levels and the dependency ratio? (Appendix 1, columns G–H.)
11. How convinced are you by the argument arising from the correlation of friendly society membership numbers with spending on the poor and pauper numbers?
12. The exercise relating friendly society membership to regional wage levels (Table 7 and Appendix 1, columns M–N) reveals nothing of significance. Why might this be the case?
13. What support exists for the 'surplus earnings' analysis (p. 505 and Table 7)?
14. In Appendix 1 the significance-test critical values are given. Explain these.

Wrigley, E.A., 'Family limitation in pre-industrial England', *Economic History Review*, 19 (1), (1966), pp. 82–109.

This essay is not recent research but it has been retained as an exercise from the first edition because it is a timeless classic and still much debated, partly because of the way in which the statistics and family reconstitution are used to support arguments about birth spacing and coital abstention as conscious forms of family limitation in pre-industrial England.

1. What is the aim of the article and why is family reconstitution necessary?
2. Exactly what is family reconstitution and what are the pitfalls of undertaking this with English parish registers?
3. Why is a nine-year moving average used in Figure 1?
4. What is meant by a 'marked inverse correlation' (p. 85)?
5. Why does Wrigley give the mean, the median and the modal age of marriage in Table 1?
6. What is age-specific marital fertility?
7. Why does the concavity of some of the curves in Figure 2 on the upper side possibly indicate family limitation?
8. What statistical procedure is used to consider whether there is a significant difference in age-specific marital fertility in the higher age groups between the early and later seventeenth century?
9. What is the point and what are the pitfalls of studying birth spacing?
10. What considerations should one bear in mind when assessing the validity of findings in Figure 4?
11. What makes Wrigley so sure that he has found evidence of family limitation? (See especially p. 95.)
12. What is the significance of Figure 5, that is, what does it purport to show?
13. Why has a chi-squared test been carried out?
14. What other explanations might there be for the statistical results?
15. What wider implications does Wrigley suggest that his findings for Colyton might have for English demographic history?

Nicholas, Stephen and Deborah Oxley, 'The living standards of women during the Industrial Revolution', *Economic History Review*, 46 (4), (1993), pp. 723–749.

1. What does the article set out to do?
2. What problems occur with using military recruitment data for heights which do not occur with using the convict data?

3. What is the normal distribution referred to on page 725 and what is seen as responsible for the heaping in the distributions of height for English and rural Irish females.
4. How do the authors compare the representativeness of female convicts, in terms of occupations and skills, with the female population of England and Ireland as a whole? What difficulties arose with this?
5. Is it a problem for this analysis that the age structure of the convicts was concentrated in the 21–31 range?
6. What weaknesses do you think there may be in the literacy data?
7. What are the three components of height-for-age? What evidence suggests that height-for-age is a good indication of nutritional and environmental impacts during the growing years?
8. Explain what is meant by the statement that urban/rural terminal height differentials were statistically significant at the 0.05 level (p. 733).
9. Why was the urban/rural height differential less important in Ireland?
10. What is a moving average and why has one been used in Figure 2?
11. What interesting thesis about the Irish potato famine does the data on male and female convict heights support? (p. 735)
12. What reasons do the authors give for the probable deterioration of English diets for women, especially in the towns, from the end of the eighteenth century?
13. What might account for the greater decline in female than in male heights over time?
14. Explain the purpose of the regression model for the composition effects by occupation (pp. 742–743).
15. From the regional regression models on pp. 743–746 what is suggested:
 - (a) about the differences in living standards between regions in England?
 - (b) about the differences between the Irish and the English economies and their deployment of household and family labour?
16. How convinced are you about the conclusions of this article given the sources and methods used?

Lindert, Peter H. and Jeffrey G. Williamson, 'English workers' living standards during the Industrial Revolution', *Economic History Review*, 36 (1), (1983), pp. 1–25.

A landmark article in its time, this piece pioneered serious use of composite and real indices in studying the movement of workers' living standards in

History by Numbers

England during the Industrial Revolution. It does however provide much scope for debating how to define and measure living standards, the representative nature of the sources and the reliability of the indices composed for this purpose. It is best read together with Feinstein's later study on the same subject. (Feinstein, Charles H., 'Pessimism perpetuated: real wages and the standard of living in Britain during and after the Industrial Revolution', *Journal of Economic History*, 58 (3), (1998), pp. 625–658.)

1. What are the aims of both of these studies?
2. What is an index? Why are indices useful? What is a composite index and what is a real index?
3. What are the pitfalls in building up a wage series for the mass of the population in this period?
4. In what way does Lindert and Williamson's wage series claim to be better than earlier ones?
5. Are there any problems with the way in which wages in the service sector are calculated and included by Lindert and Williamson?
6. Does the Lindert and Williamson new wage series determine the conclusions of the article?
7. What do Lindert and Williamson say is wrong with earlier price series?
8. How is the price series for cotton derived by Lindert and Williamson?
9. What other alterations are made in the price series by Lindert and Williamson compared to earlier price series?
10. Have you any comments about the incorporation of rents by Lindert and Williamson?
11. Why are weights so important and yet so problematic in constructing price series?
12. What do Lindert and Williamson say about unemployment and how do they allow for it? Is this satisfactory?
13. How do Lindert and Williamson allow for family earnings?
14. Is family size important?
15. What other forms of income should one include in ascertaining well-being?
16. Should lifetime income be considered?
17. What are benchmark years?
18. What are disamenities and how are they measured?
19. What is the importance of considering push factors in migration in relation to the disamenities argument?

20. Compare the Lindert and Williamson article with that of Charles Feinstein (detailed above). What are the main differences? What improvements has Feinstein introduced? What further evidence would be welcome?

Allen, Robert C., Jean-Pascal Bassino, Debin Ma, Christine Moll-Murata and Jan L. Van Zanden, 'Wages, prices and living standards in China 1738–1925: in comparison with Europe, Japan and India', *Economic History Review*, 64, (2011), pp. 8–38.

1. Why is the comparative evidence of global wages and prices described as 'fragile'?
2. How is the nominal wage series for China constructed and what are the likely pitfalls?
3. What is the purpose of the wage regressions featured in Table 2?
4. Why have the nominal wage series for China and Europe been converted into grams of silver and why might this be questioned?
5. What are 'baskets of goods' and why must they be different for Europe compared with China for an exercise such as this?
6. Why might one question the basket of goods distribution for Europe as a whole (or for China as a whole)?
7. What problems might attend the application of the price series to the baskets of goods?
8. What are the main pitfalls in estimating any real wage measure?
9. What is the likely effect of applying the silver standard to real wage measures?
10. Why is it a concern that the European and Beijing baskets employ Paasche and Laspeyres price indices?
11. What is the Fisher Ideal Price Index and why is it used here?
12. Why is it seen to be necessary to compare a 'bare bones' level of subsistence and a 'respectable' level?
13. What are welfare ratios and how are these helpful?
14. Describe what Figure 6 purports to demonstrate.

Oxley, Deborah and David Meredith, 'Food and fodder: feeding England, 1700–1900', *Past & Present*, 222 (2014), pp. 163–214.

1. What is the main question posed in this article?
2. What is energy cost accounting?
3. Why are the three existing accounts of calories available per capita so different?
4. Write a short account of the findings available in Figure 2.

History by Numbers

5. What do anthropometrics add to the debate?
6. Discuss the efficiency of the graphs in Figure 3 in conveying various elements of information.
7. How well does the height and energy accounting data fit together in Figure 4?
8. Discuss the advantages of dividing the population under observation into deciles in Table 3 and elsewhere.
9. How reliable are the ‘models of egalitarianism’ used and how might their reliability affect the argument?
10. Discuss the complementary information provided by dietaries and the use of Figure 5 to show this.
11. To what extent does Figure 6 reconcile the various views, estimates and evidence into a coherent picture?
12. How close did Britain come to a Malthusian crisis in the nineteenth century?

Benjamin, Daniel and Levis A. Kochin, ‘Searching for an explanation of unemployment in interwar Britain’, *Journal of Political Economy*, 87 (3), (1979), pp. 441–478.

This was a controversial article when it first appeared as it was used to bolster Thatcherite policy with respect to unemployment benefits. It includes a rather crude model and time series analysis. It has been retained from the first edition as a good example of the pitfalls of these methods.

1. What does the article set out to do and what are the ‘three solid strands of evidence’ upon which the argument is based?
2. What is time series evidence? (p. 442)
3. Describe the graph which is in Figure 1 and explain what it shows.
4. What three aspects of the benefits system contributed, in the view of the authors, to easy movement from employment to unemployment?
5. What sort of unemployment are Benjamin and Kochin seeking to explain and why have they little interest in unemployment that might arise from deficient demand?
6. What ratio is the basis of their time series analysis?
7. From what sources are B, W and U calculated? Are you satisfied with the reliability of the sources for these purposes? If not why not?
8. What is reverse causation and how do the authors cover themselves against this being the reason for their time series findings? (p. 455)
9. How does the behaviour of juvenile unemployment appear to add weight to their argument?

10. What other reasons might there be for such movements in the rates of juvenile unemployment?
11. What weaknesses are there in the aspect of their argument that relies on the movement of unemployment rates of married women?
12. Especially in view of the calculations attempted in the final section, what is your view of the relationship between the sources and data and the quantitative methods attempted in this research?

Botticini, Mariestella, 'A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town, 1415–1436', *Journal of Economic History*, 59 (1), (1999), pp. 104–121.

1. What is a dowry?
2. List the potential implications of dowry payments as (a) a marriage payment and (b) an intergenerational transfer.
3. What sorts of other things may need to be held constant by applying the *ceteris paribus* notion on p. 106?
4. Comment on the size and nature of the sample of marriage contracts.
5. What sort of information can be derived just from Table 1, assuming the marriage sample was representative of Tuscany as a whole? Look at the difference between medians and means here, and at the measures of dispersion.
6. What relationship appears to be present between the male–female age gap in marriage and the prevalence of women 'marrying down'?
7. What is the 'present net value hypothesis'? (pp. 106 and 109–121) With what justification is this tested by running a correlation test upon dowry size and bride's age?
8. What is meant by the altruism model and how is it proposed that it should be tested?
9. What is the purpose of Table 4 and does it succeed?
10. Has the problem of measuring the wealth of groom and bride households been satisfactorily solved for cases where the *Castato* evidence postdates the marriage contract?
11. What are dummy variables? Comment on their use here.
12. In Table 5 what is meant by:
 - a. the *t*-statistic?
 - b. *R* squared (R^2)
 - c. The levels of significance indicated by the asterisks?

History by Numbers

13. Is the finding that dowry size and fertility were correlated of any causal significance? (p. 117)
14. To what extent are the present net value and altruism models supported by this research? What other unrelated explanations could help to account for these findings?
15. Given the sex ratios in Table 1 in the appendix, what is wrong with calling Cortona a ‘Tuscan town’ in the title of the article? How would factors peculiar to Cortona work to reduce confidence in the applicability of these research findings to the rest of Tuscany?

Oxley, Deborah, “Pitted but not pitied” or, does smallpox make you small?’, *Economic History Review*, 59 (3), (2006), pp. 617–635.

This article is the end point (for the time being) of a debate about whether smallpox stunted growth in the eighteenth and nineteenth centuries. It is an example (as is Nicholas and Oxley above) of anthropometric history (that is, using height or other human data to infer trends in standards of living, health or nutrition).

In order to do this exercise it is wise to have read the earlier exchanges:

Voth, Hans-Joachim and Timothy Leunig, ‘Did smallpox reduce height? Stature and the standard of living in London 1770–1873’, *Economic History Review*, 49 (3), (1996), pp. 541–560.

Oxley, Deborah, ‘“The Seat of Death and Terror”: urbanization, stunting and smallpox’, *Economic History Review*, 56 (4), (2003), pp. 623–656.

Leunig, Timothy and Hans-Joachim Voth, ‘Comment on Oxley’s “Seat of Death and Terror”’, *Economic History Review*, 59 (3), (2006), pp. 607–616.

1. Why is it impossible to use the anthropometric data to test for the stunting impact of diseases other than smallpox?
2. Comment on the assertion that larger sample size has made London appear unique in Oxley’s earlier publication (2003).
3. Is stunting an independent or a dependent variable? Do the authors agree on this?
4. What do the 95 per cent confidence intervals tell us in Figure 1?
5. What is meant by meta-analysis?
6. How is the line in Figure 2 affected by outliers?
7. Why is Oxley so against aggregations in the regression of height against pock marking?
8. What is the effect on the findings of increasing the sample size (Figure 3)?
9. What differences between the nature of the convict sample and the prisoner sample should investigators of this question be wary about?

10. How clear and convincing do you find the argument about statistical significance on p. 627?
11. What are the p -statistics quoted at the bottom of p. 629?
12. Explain: ‘the relationship between smallpox and stunting appears to be mediated through location’ (p. 631).
13. What final exercise does Oxley carry out (pp. 631–632) to strengthen/‘prove’ her point about the importance of urban disamenities and that they may have been the key independent variable?

Gutmann, Myron P., Sara Pullum-Piñón and Thomas W. Pullum, ‘Three eras of young adult home leaving in twentieth-century America’, *Journal of Social History*, 35 (3), (2002), pp. 533–576.

1. Discuss the advantages and the limitations of using the census for analysing the age at leaving home.
2. Using the technical terms you have learned in this book, describe Figure 1: how it was derived and what it shows.
3. Considering Table B1, clearly in some periods and for some groups the age of leaving home was more dispersed. Can you spot and comment upon the likely cause of these variations? (Table B8 might be useful here.)
4. List the range of causal variables likely to influence the age of leaving home.
5. Compare your answer to Question 4 with those causal variables likely to determine the stayers in the parental home. Are they the same?
6. Define Logit analysis and describe how it is used in this article.
7. Comment on the lack of socio-economic information of the parents in the census for this sort of analysis.
8. How has the likelihood of people declaring their occupation in the census changed before and after 1910?
9. Why are changes in the percentage ever married inversely correlated with the age of marriage (p. 536)? In what circumstances might this not apply?
10. Explain the characteristics of married young people still living with parent(s) based on the figures in Table B9.
11. What are the relationships between the results shown in various figures and tables and the results of logistic regressions shown in Appendix C?

Etro, Federico and Laura Pagani, ‘The market for paintings in Italy during the seventeenth century’, *Journal of Economic History*, 72 (2), (2012), pp. 423–447.

1. What are the advantages and the pitfalls of using the art market and the price of paintings in this period in order to view more closely the operation of supply and demand and of ‘rational behaviour’?

History by Numbers

2. Define moral hazard and explain why it was common in such transactions.
3. Why are the markets for northern and central Italy examined separately?
4. List the supply and demand side factors that the authors suggest may have influenced the price of paintings.
5. List all of the variables obtainable from the evidence in the contracts collected by Spear and Sohm.
6. What drawbacks might attend the use of evidence collected in this way by other researchers?
7. Which three further variables do Etro and Pagani gather?
8. How well can the perceived quality of the painters be proxied by the artists' fixed effects?
9. Comment on the meaning of the means and standard deviations shown in Tables 1 and 3, including those for dichotomous variables.
10. Comment upon the findings in Tables 2 and 4. Are you convinced by them?
11. What is a dummy variable? How are these used in the analysis of the 'destination effects' in the Venetian market (pp. 437–438), and in the analysis of positioning effects in the central Italian market (p. 440)?
12. What does Figure 1 show?
13. How convinced are you by the age/price relationship purported to obtain with 'conceptual' innovators?
14. To what extent are you convinced by the overall argument that the techniques of regression on these data allow one to demonstrate the operation of modern market behaviour?

Dennison, Tracy and Sheilagh Ogilvie, 'Does the European Marriage Pattern explain economic growth?', *Journal of Economic History*, 74 (3), (2014), pp. 651–693.

1. What is the European Marriage Pattern (EMP)?
2. Why is the EMP generally linked in a positive way to economic growth?
3. How do the authors intend to ensure that the apparent country differences do not merely reflect biased availability of data for different societies at different periods?
4. What potential drawback attends only the use of published research findings in this composite dataset?
5. How confident are you that the authors take sufficient care to account for the bias that may enter demographic results because of different ways of recording and calculating vital events or demographic/household characteristics across different cultures and time periods?

6. What are the drawbacks of using GDP per capita estimates as the measure of economic growth or economic success for this research?
7. Describe the nature of the regression results laid out in Table 2 and their reliability.
8. What variations in female lifetime celibacy within Europe does the study highlight?
9. What difficulties attend the examination of the household complexity element in the EMP study?
10. What is a Borda ranking and what do the Borda rankings featured in Table 5 add to the analysis?
11. Why did the social and economic position of women vary independently of the EMP?
12. Discuss the role of the EMP compared with other possible factors affecting human capital formation in Europe in the eighteenth and nineteenth centuries.
13. To what extent does the EMP explain demographic responsiveness to economic conditions?
14. Why is the wider cultural environment likely to have been important?
15. What non-familial institutions are likely to have influenced female age of marriage, celibacy and household structures, and in what ways?

Olney, Martha J., 'When your word is not enough: race, collateral and household credit', *Journal of Economic History*, 58 (2), (1998), pp. 408–431.

This is a more complex article but it is very clearly expressed and has a good blend of social and economic history, the application of some economic theory, as well as multiple regressions.

1. Discuss the potential pitfalls of using the 1918–1919 Consumer Purchases Survey to consider credit and race, given the way in which it was organized and executed (pp. 409–410).
2. List all of the reasons (in theory) why black families might have:
 - used less shop or merchant credit than whites;
 - used more instalment credit than whites;
 - had higher rates of savings than whites on similar incomes.
3. Are there any ways of summarizing and simplifying the data contained in Table 1 that would strengthen the case made in the text?
4. Do you have any further comments to make about the variations of instalment credit related to different items in Table 3?

History by Numbers

5. Discuss the impact of chain stores upon grocery credit.
6. Why are small sample sizes a consideration on p. 418?
7. Have the problems that you identified in relation to Question 1 been addressed on pages 417–418? List those problems that you feel still remain.
8. List the independent variables in the logit analysis on the supply side and on the demand side, and check whether you can identify these in column 1 of Table 7.
9. Rehearse the aims and potential pitfalls of multiple regression analysis.
10. What are the 90 and 99 per cent confidence indicators mentioned in the notes to Table 7?
11. What are the *t*-statistics referred to in the notes to Table 7?
12. Do you see any problems in the proxy for wealth described in the notes to Table 7?
13. Define ‘asymmetric information’ and discuss the use of this concept in the argument on pages 424–427.
14. Do you find the argument about poor black families substituting savings for merchant credit convincing?
15. Explain in simple terminology: ‘The creditor remedy . . . offset the risks of adverse selection and moral hazard?’

CHAPTER 8

ECONOMIC HISTORY AND ECONOMETRIC HISTORY

The branch of history most closely associated with the notion of ‘history by numbers’ is economic history. Interest in economic change and economic growth has always necessarily involved an interest in quantities and measurements – of population growth, agricultural and industrial outputs, raw material inputs, exports, imports, prices, wages, and productivity. This interest was endorsed during the inter-war and post-war years of the twentieth century when inequalities of the global growth process became increasingly apparent and when most Western governments became more involved in state intervention to aid the functioning of market economies, as a result of the Keynesian revolution. These developments gave a spur to the generation of economic and social statistics – building in Britain upon earlier traditions of political economy and the Victorian statistical movement.¹ Twentieth-century developments in economic theory, policy and planning also influenced the ways in which historians approach the past giving them new analytical tools that can be applied retrospectively. These developments form the context of the rise of quantitative historical analysis of the economy that accompanied the expansion of economic history as a discipline in universities, in many parts of the world, in the third quarter of the twentieth century.²

Once historical statistics relating to economic matters are collected, and even in the collecting of them, ideas derived from economic theory are inevitably brought into play. As Deirdre McCloskey has argued: ‘little of what historical economists do by way of collecting statistics escapes the touch of economic theory’³ Equally, as Herbert Spencer originally pointed out: ‘It is a truth perpetually that accumulated facts, lying in disorder begin to assume some order when a hypothesis is thrown among them’.⁴ These two quotations to some extent highlight the potential but also some of the pitfalls of much quantitative economic history.

The combination of modern economic theory with quantitative historical data is an approach that has become increasingly prominent in the second half of the twentieth century, particularly in those countries (the USA is the most prominent) where economic history has developed largely within, and allied to, economics faculties. One only has to look at the nature of articles published in the *Journal of Economic History* (the main US outlet for the subject) and the *Economic History Review* (the main British publication), to see how important statistical techniques have become in the last 35 years or so.⁵ The US journal *Explorations in Economic History* has for long been exclusively devoted to publications arising from this approach.⁶ In such articles statistics are usually, very deliberately, allied to the testing of formal (mathematical or statistical) models

History by Numbers

derived from economic theory. This sort of history experienced an initial boom in the late 1960s, 1970s and early 1980s spawning a great deal of criticism of the impact that it had upon historical understanding, as was discussed in Chapter 2.⁷ In this phase most research of this kind rested on the application of simple price theory equations with all of the free market and ‘efficient information flow’ assumptions that such modern (neoclassical) economic analysis of markets accepts for the purposes of simplification. It was applied to a relatively restricted range of historical questions mostly concerning the private and public benefits of particular institutions (such as slavery) and of particular technological innovations (such as railways and steam ships). Quantitative economic history has subsequently developed with a much wider range of more sophisticated tools, assumptions and models, and with a growing variety of topics of enquiry: from research on living standards and nutrition using anthropometric data to the micro-analysis of decision-making in firms and households. Growth theory and the analysis of long-term economic growth (often comparative) is now at the forefront of much quantitative economic history, and more attention is paid to the modelling of institutions and information. Such relatively advanced forms of quantitative analysis remain the province of only a subset of economic historians but their methods and conclusions are more readily accepted and less often debated than in the past partly, but not only, because the quantitative techniques applied are poorly understood by the historical mainstream.

Quantitative economic history thus provides a useful case study to illustrate some of the most fundamental issues concerning history by numbers more generally, especially where this is wedded to economic (or other) quantitative modelling. Economic historians will no doubt find this chapter a good deal more interesting and useful than other historians but many of the debates raised by economic history have a wider resonance and relevance to users of quantitative methods in history as a whole.

Some definitions

Econometrics is the application of mathematical statistics to economics. During the last half-century academic economics has come to be dominated by this approach. It generally involves the building of formal economic models and their statistical testing.⁸ Econometrics is the core component in the training of an economist, and econometric analysis is regarded as the high status, cutting edge of virtually all applied research in economics.⁹ Fifty years ago, before the so-called ‘formal revolution’ this was much less the case.

Econometric history is the application of modern economic theory and methods (including the centrally important statistical methods, regression in particular) to history. Most of the statistical methods used are simple and have already been covered in this volume. Descriptive tools such as graphs, frequency distributions, indices, growth rates and trends are joined by inferential statistics involving correlation and regression analysis, sampling and significance testing. In econometric history these are used in combination with economic theory and model building particularly in supply and

demand arguments, general and partial equilibrium theory, business cycle analysis, national accounts (input/output) research, work on comparative advantage, research on health and nutrition based on anthropometric evidence. These models are normally informed by and expressed in graphs, multiple regression analysis or in algebra, forming difference equations or functional equations (see below). Econometric history has been described as being 'born of the marriage contracted between historical problems and advanced statistical analysis, with economic theory as bridesmaid and the computer as best man'.¹⁰

Cliometrics refers to the application of statistical methods to historical facts in the interest of history. This often includes much purely descriptive statistics but to the extent that economic theory is based upon inferential statistics, there is clearly an overlap between cliometrics and econometric history: the two terms are often used interchangeably. Indeed a large part of what is referred to as econometric history, especially as it was practised in the first wave of 'New Economic History' in the 1960s and 1970s, was really just cliometrics (the formal modelling element based upon economic theory was by no means always present). Furthermore, as the theory and method of economics has, in recent decades, become more oriented to statistical theory, its application in history has increasingly fitted the cliometric rather than the econometric label.¹¹

Economic **model building** in history refers to the construction of models of historical economies or sectors. Like most econometrics these have been almost exclusively based upon the ideas and assumptions of neoclassical economics and, to a lesser extent, upon new and older forms of classical economics. Occasionally, Marxist models or other alternative models have also been employed. Usually data is incorporated into a model of the interaction of variables in an economy or sector. The model is then used to simulate (in a simplified way) the operation of the major influences so that unknown elements in the model can be estimated or the model can be employed to understand events and circumstances in the past. Many models employ regression analysis including multiple regression techniques in investigations of cause and effect. Most models employ inferential statistics of some kind because their aim is to manipulate the data available in order to infer a great deal more about missing variables or about the functioning of the economy than would otherwise be possible.

Extrapolation is often used to estimate missing values on the basis of the existing data. Often this is derived from extending data that displays a definite linear trend or curve in time series or to introduce predictive power into a regression line. The technique of **back-projection** is a form of extrapolation used to infer or estimate figures for periods where the data is unavailable, on the basis of later periods where figures do exist. For example Wrigley and Schofield, with the assistance of the demographer R. D. Lee, used a back-projection of population data from the 1851 census into the eighteenth century in order to calculate the shortfall of vital events, and hence population, recorded in parish registers. In this case the back-projection is based on a simulation of demographic increase based on known or estimated facts about family size, marriage ages and life expectancy.¹²

History by Numbers

A strong current in econometric history especially in the 1960s and 1970s was **counterfactual history**. This is the calculation of costs and benefits of a particular innovation or institution in the past compared with the costs and benefits that would have been obtained in the absence of that innovation and using a 'second best' system. The difference in total net cost to society is the **social savings** gained or lost.¹³ The social savings gained by having a railway system, for example, is the cost of the economic resources that would have been required in the second best transport system and which can then be released for alternative uses. Counterfactual history is based on the notion that we cannot tell how much an event or innovation matters in the past without calculating what would have happened without it. This idea is not unique to econometric history: historians have commonly asked 'what if' sorts of questions and there has been something of a tradition of qualitative counterfactual studies sometimes referred to as **virtual history**.¹⁴

The history of econometric history

Econometric analysis has an interesting history in the context of different national traditions and through the activities of particular exponents. This history, up to the 1950s, has been the subject of intensive study.¹⁵ The development of econometrics was delayed until well into the twentieth century partly because most economists of the nineteenth century believed that mathematics and statistics worked in different ways: maths as a tool of deduction and statistics as a tool of induction. Statistical thinking had been incorporated into economic writings in Britain from the eighteenth century, particularly in measurements of economic variables, as we have seen. Such descriptive uses of statistics remained strong amongst economists of the historical and institutional traditions in Europe and the inductive approach was strengthened by a conscious rejection of what was seen as the increasingly deductive, abstract, and ahistorical neoclassicism of the marginal school of economics in the late nineteenth century.¹⁶ In neoclassical economic analysis the adoption of statistical tools and the analysis of large bodies of data had to wait until statistical theory had become a more sophisticated tool of *deductive* analysis. Even then it was often the case that economists and historians who used statistical *evidence* were often those who most strongly rejected mathematical models and methods.¹⁷

In the 1930s the international Econometric Society grew in influence and the journal *Econometrica* was founded which included economists from many different schools of thought. What supporters of econometrics in this period had in common was not so much adherence to just one sort of theory but the desire to conjoin mathematical economics and statistical economics.¹⁸ By the 1940s this project had become firmly established and econometrics has been the most important form of applied economics ever since. Econometrics became dominant as a mode of enquiry and analysis in economics with the new formalism (mathematical model-building and testing) of the 1950s and 1960s and it then grew more prominent in historical research particularly in

the United States. Because, in the USA, economic history in universities had always been closely tied to economics departments, it was natural that when econometric techniques became more prominent in economics, they would also feed through into historical work there in particular.

During the late 1960s and 1970s the early econometric studies in history came to be referred to and seen as the New Economic History. Famous studies appeared on European manorialism, transport (especially, railway innovation), the economics of slavery, open field agriculture in England, the quality of nineteenth-century entrepreneurs, British industrial competitiveness, and the causes of the Great Depression in 1930s America.¹⁹ During this period its most vigorous exponents claimed it would eventually provide *definitive answers* to many of the most fundamental questions asked by economic historians. The implication was that this approach could restore objectivity and dispassionate scientific precision to history.²⁰ In short, the econometric movement was underpinned by a renewed positivism (which was to create much critical reaction).

Peter Temin described econometric history in its heyday in 1973 as 'a form of applied neo-classical economics':

Examples ... typically start with a formal model of some aspect of economic behaviour, assemble data for use in the model, and draw conclusions by joining the data and the model. The last step can be done in many ways: by constructing hypothetical answers to questions under varying assumptions, by estimating parameters to specified equations, by using facts in the context of a deductive argument to reject alternatives. The common element is the use of a specific model with explicit assumptions.²¹

At the time he was writing, in the first wave of econometric history, most formal models in use were designed to test whether the market for a certain good was working effectively or not (that is, whether equilibrium was obtained via the price mechanism), and models were often framed in such a tautological manner that they inevitably proved that 'The market, God bless it, works'.²² Indeed, beyond that, it has been suggested that the first wave of so-called cliometrics had only a limited connection to econometrics as used in economics at the time. Nevertheless, particularly in the USA, the early practitioners of cliometrics, by explicitly applying the first principles of economics, considerably contributed to historical debates concerning economic growth, trade, capital markets, some technological innovations, industrialization, westward expansion, slavery and economic fluctuations.

The econometric tools that historians of the 1960s and 1970s were using were relatively unsophisticated compared with today yet little effort was made to address the technical limits to economic theorizing. The underlying assumptions of the models were oversimplified and insufficiently related to the context in which they were being applied. Little attention was paid to institutions or to cultural or social variables that lay outside the models and could not be easily quantified. Enthusiasm for modelling and the testing of models often displaced critical discussion of the reliability or detail of the historical

History by Numbers

sources themselves. For all these reasons but also because of the deep conservatism of historians, only a restricted number became converts to the 'New Economic History' and initial enthusiasm tended to die away in the 1980s hastened by the criticisms of some economists like Solow:

As I inspect some current work in economic history, I have the sinking feeling that a lot of it looks exactly the kind of economic analysis I have just finished caricaturing: the same integrals, the same regressions, the same substitution of t-ratios for thought.²³

The boom in econometric history in the 1960s and 1970s is often seen today as representing a 'first wave' whose achievements came far from matching their ambitions and whose work provoked a negative reaction from which cliometrics is still struggling to recover. This is certainly the case in Britain.

Econometric history: first wave examples

The energy and enthusiasm which went into this first wave of econometric history, the results that it produced and the criticisms that it sustained are important for anyone hoping to understand the nature of 'history by numbers'. Let us take some important examples.

Railway history

A very significant cluster of studies in the 1960s and 1970s were concerned with railway innovation. These studies formed a major experiment in quantitative counterfactual history. They took the form of cost–benefit analyses applied to the past. They were concerned to calculate **social savings**, that is, the extra cost of the economic resources which would have been required in using the second best transport system. The path-breaking work here was R. W. Fogel's *Railways and American Economic Growth: Essays in Econometric History* (1964). In this Fogel constructed a model of what the US economy might have looked like in 1890 if railways had never existed. He chose 1890 as a year by which many of the benefits of the railways would have been apparent but otherwise the choice is arbitrary. He justified the need for such a benchmark because it was easier to calculate social savings at a finite point in time and because there was less scope for error than would have been the case if social savings had been calculated over a longer time period. The major problem with Fogel's study, and with all similar calculations, is that the construction of the hypothetical/counterfactual economy involves many arbitrary and subjective decisions. Fogel, for example, allowed his hypothetical US economy to respond in the absence of railways and he created an imaginary canal system and road system designed to carry the same volume of goods between the same destinations as did the rail system.

Fishlow's study of the same subject not only took an earlier benchmark of 1859 but did not allow the hypothetical economy to react to the absence of railways because of the risk of introducing too many additional errors into the calculation.²⁴ The trouble is that the counterfactual world must necessarily involve adjustments otherwise it is of limited use to social savings calculations. But in creating a hypothetical world the door is open to a multitude of different alternatives. A close look at Fogel's work highlights some wider problems in constructing the counterfactual model. He can for example be criticized because goods in this hypothetical scenario would not necessarily travel between the same destinations in the absence of railways. He also assumes that railways do not create traffic and that higher-cost transport does not lower demand (in economists' parlance he assumes that the price elasticity of demand for transport is low). He also assumes that railways and alternative forms of transport are perfect substitutes whereas, for many goods they clearly would not be. Fogel defends his approach to some extent by saying that he does not have to create an entirely accurate counterfactual model in order to show that the social savings of railways were relatively small. His aim from the outset was not accurately to calculate social savings but to question the long-held assumption (based largely on impressionistic evidence) that railways were indispensable for the American economy. To do this all that was necessary, he argued, was to make sure that his hypothetical world was one which would allow social savings estimates for railways to be maximized. Where a range of estimates was possible in his calculation he always erred toward biasing the case against himself and in favour of the impact of the railroads. Thus he showed that even with the most generous of his estimates, the role of railways in American economic growth had been previously overstated to a considerable degree.

Emulatory studies of the railways in other national contexts came to similar conclusions: that railways speeded up economic growth but that the same growth could have been achieved by alternative forms of transport. Several studies also attempted to place a precise figure upon GNP savings to the economy of rail development. G. R. Hawke's study of England and Wales, for example suggested that there had been a saving of 4 per cent GNP (£28 million) whilst Metzer's study of Russia placed the figure there at 4.6 per cent. Hawke and Metzer used the same approach as Fogel: they both took a benchmark year for measurement (1865 and 1904 respectively) and both allowed their hypothetical economies to adjust to the absence of railways. But in attributing a figure to social savings and making this the centre point of their research they drew more criticism than Fogel had because of the difficulties involved in making the counterfactual model accurate.²⁵ Calculations for America, Russia and other large, relatively landlocked and agrarian economies were criticized, in particular, because of the difficulties in estimating the wider benefits brought about by such innovation. In relatively backward economies where alternative forms of transport are not viable, railway innovation considerably effects costs and prices throughout the economy, opens up huge new areas to trade and makes it impossible to estimate any road/canal alternatives.²⁶

All such studies of course suffer from the same problems that dog the technique of cost-benefit analysis as an aid to economic decision-making in the present. As everything has to be costed in monetary terms, decisions have to be made about the value to be

History by Numbers

placed on relatively intangible factors such as health, pollution, wider travel and economic horizons, if these are to be included. The problem of where to draw the line with such analyses, what to include and what to omit, is also crucial. When does one decide that all significant costs and benefits have been estimated and accounted for? How does one decide what is significant? Do we include indirect costs and benefits such as the effects of railways not just upon the cheaper and more efficient supply of perishable foodstuffs but, through this, upon nutrition and efficiency of the workforce? A third difficulty concerns uncertainties. How do we decide what will be the price of coal or other important commodities under different hypothetical economic circumstances? And do market prices in any case represent the value placed upon commodities by society?

Construction of the counterfactual world without railways is obviously highly subjective with the difficulty in particular in deciding how economic activity responds to the absence of railways. Do the same goods pass between the same destinations using roads, sea or canals or must the model allow for the development of different sectors and different sales and freight flow patterns? How would other forms of transport have developed in the absence of railways? There is much scope for ignorance and guess work.²⁷

Slavery

Conrad and Meyer's work on slavery is often regarded as marking the beginnings of the 'New Economic History'. They postulated that American 'negro slavery' was characterized by two production functions and that an efficient system developed whereby those regions best suited to the production of cotton specialized in agriculture whilst less productive areas produced slaves to be exported to the staple crop areas. The model applied was a neoclassical two region, two commodity trade system. The data required for this path-breaking exercise were slave prices, cotton prices, average outputs of field hands and field 'wenches', the life expectancy of persons born in slavery, the cost of maintaining slaves during infancy and other non-productive periods and the net reproduction rate and demographic composition of the slave population in the 'breeding' and using areas. Table 8.1, one of many produced by their research, gives a flavour of their work: the quality but at the same time the cold and inhuman language used which is partly what drew the fire of critics. Conrad and Meyer concluded, in the face of other opinions, that slavery was profitable for all the South at the time of the Civil War and that political forces were required if it was to end.

Robert Fogel and Stanley Engerman developed this sort of approach with their landmark study of the profitability of American slavery: *Time on the Cross: The Economics of American Negro Slavery* (1974).²⁸ This used statistics from plantation records and censuses to demonstrate that planters in the mid-nineteenth century were a rational and humane group and that slaves were prosperous and well-treated. They also confirmed that slavery had not ceased to be profitable to owners at the time of the Civil War. This work led to a most acrimonious debate with historians who suggested that *a priori* prejudices

Table 8.1 Annual returns on a ‘prime field wench investment’ (working on land which yielded 3.75 bales per prime male field hand, assuming a 7.5 cent net farm price for cotton and 10 ‘saleable’ children born to every wench)

Year from purchase date	Personal field returns (\$)	Child field returns (\$)	Child sale returns (\$)	Personal upkeep (\$)	Child upkeep (\$)	Net Returns (\$)
1	56			20		36
2	40			20	50	-30
3	56			20	10	26
4	40			20	60	-40
5	56			20	20	16
6	40			20	70	-50
7	56			20	30	6
8	40	3.75		20	80	-56.25
9	56	7.50		20	45	-1.50
10	40	15.00		20	95	-50.00
11	56	22.50		20	60	-1.50
12	40	37.50		20	110	-52.50
13	56	52.50		20	75	13.50
14	40	75.00		20	130	-35.00
15	56	97.50		20	95	47.50
16	40	127.50		20	150	-2.50
17	56	157.50		20	115	78.50
18	40	195.00		20	165	55.00
19	56	232.00		20	130	134.30
20	40	195.00		20	170	920.00
21	56	232.50		20	130	138.00
22	56	195.00	875	20	120	986.00
23	56	232.50		20	120	148.50
24	56	195.00	875	20	110	996.00
25	56	232.50		20	110	158.00
26	56	195.00	875	20	110	1,006.00
27	56	232.50		20	100	168.00
28	56	187.50	875	20	90	1,008.50
29	56	225.00		20	90	171.00
30	56	180.00	875	20	80	1,011.00
31		210.00			80	130.00
32		157.00	875		60	972.50
33		180.00			60	120.00
34		120.00	875		40	955.00
35		135.00			40	95.00
36		67.50	875		20	922.50
37		75.00			20	55.00
38			875			875.00

Source: A. H. Conrad and J. R. Meyer, ‘The economies of slavery in the antebellum South’, p. 63 in A. H. Conrad and J. R. Meyer, *Studies in Economic History* (London 1965), pp. 43–72; article first published in *Journal of Political Economy* (April 1958).

History by Numbers

had led to selective use of evidence and a rosy view of both slavery and the motives of slave owners. A major criticism was that Fogel and Engerman had misinterpreted slavery because of a desire to make everything fit into a neoclassical model in which 'each and every slave owner regarded slaves solely as productive instruments and used them for a single transcendent purpose: the maximization of pecuniary gain'.²⁹

Economic growth and entrepreneurship

Econometric debate about the performance of the British economy and of British entrepreneurs in the late nineteenth century was ignited by McCloskey's 1970 article 'Did Victorian Britain fail?'³⁰ This followed a considerable traditionally researched literature that had blamed Britain's retardation and relative economic decline upon the inefficiencies of factor and product markets, low rates of investment and the conservative behaviour and attitudes of British entrepreneurs of the period. McCloskey applied a very crude neoclassical model to argue that markets were working well and that the late Victorian economy was growing as fast as was possible. Formalizing the problem of British entrepreneurship in terms of neoclassical profit maximizing models also showed that generalizations about the conservatism and incompetence of entrepreneurs was exaggerated. Slow adoption of the basic process in steel-making and the retention of mule spinning in cotton were, for example, correct (that is, 'rational') choices given relative factor costs.³¹ These arguments drew much criticism.³²

Similarly contentious opinions were expressed about the impact of the Napoleonic Wars upon investment and growth in Britain. J. G. Williamson and others emphasized the extent to which the markets for labour and capital failed and therefore slowed down growth.³³ Williamson's later work on nineteenth-century urbanization in Britain was important in highlighting models of rural–urban migration and in arguing that underinvestment in city infrastructures had a negative impact on the labour market.³⁴ G. R. Boyer made a similar detailed analysis of the impact of the English poor relief system upon migration and economic growth during the same period using multiple regression models.³⁵ The crude modelling and the results of these pieces of work are open to question but they did pave the way for more sophisticated analysis of the measurement and operation of push and pull factors in migration and of the operation and impact of welfare measures.

One of the more audacious pieces of New Economic History was Williamson's attempt to model the entire US economy in the nineteenth century combining some counterfactual exercises with a general equilibrium model. He used 72 equations in the model but even this number in Williamson's own view did not capture much detail. To keep the calculations viable Williamson confined his analysis to the markets for labour, capital, part of the land, all manufactured goods taken together and all agricultural production. Services were not included thus three factors of production and two sectors were made to stand for all America's assets. But all models need to simplify and the test for Williamson and others is whether their particular simplification appears to yield a useful addition to knowledge. Amongst other things, he attempted to test the validity of

traditional positive interpretations of the impact of the open frontier by means of two counterfactual models: first with the frontier closed in 1870 and second with a boundless frontier as had been the case earlier in the century.³⁶ He suggested that the closed frontier would have had the effect of raising the output per acre of eastern agriculture by 120 per cent instead of the 19 per cent which it experienced, and that the open frontier slowed the shift of labour into secondary and tertiary employments. In a review of Williamson, Clive Lee argued that:

Whether the mathematics and estimates in such models are completely accurate is of secondary importance compared with ... the possibility of eliminating ... a wide range of erroneous relationships, patterns of growth, assumptions about cause and effect, and even entire explanatory constructions.³⁷

This comment encapsulates wider criticisms and evaluations of econometric history of the first wave.

Neoclassical model-building

As the dominant form of economic theorizing in the 1960s and 1970s was neoclassical, the cliometrics of the period almost always employed the methodology and assumptions of neoclassical theory. The nature of this theorizing and the sorts of economic relationships and models arising from it are important in understanding much of the criticism of econometric history, especially in its first wave. Most criticism revolves around rejection of the ideological underpinnings of free-market, laissez-faire theorizing. What is often at stake between neoclassical theorists and their critics is optimism or pessimism about the functioning of markets. In addition there is fundamental disquiet over the applicability of neoclassical theory to economies and markets in the past. It is important to address these issues here because, despite appearances to the contrary, much econometric history still incorporates the same neoclassical market-oriented assumptions as it did in the past.

At the heart of the neoclassical model is the view of the economy as a competitive regime of production and exchange where the price system allocates resources in a semi-automatic way. Study of the price system and the formation of models based on it can only be done with the aid of a range of assumptions (often called stylized facts). These assumptions, though recognized as false or exaggerated, are regarded as being sufficiently close to reality when large numbers of economic actors are involved.³⁸ The stylized facts enable certain features of the environment and of human motivation to be taken for granted (or held to be constant). This enables one to abstract the economy or economic activity from the wider context of society and social relationships so that models can be constructed with a manageable number of variables.

What are the most important assumptions? And are they affected when we consider applying neoclassical analysis to the past? There are two main postulates:

History by Numbers

- a) **The market postulate and market clearing:** this assumes that markets will achieve equilibrium and clear in the short or medium term which assumes a competitive environment, near perfect mobility of factors of production and hence good communications and information flows.
- b) **The rationality postulate:** this states that if an individual is presented with a situation of choice in an economic setting he/she will act to optimize his/her economic position.

Institutions that protect private property and enforce contracts are seen as important in providing the conditions for the market and rationality postulates to be fully developed. In short, then, most neoclassical theory is based on a logical analysis of the maximizing behaviour of large numbers of well-informed and independent individuals active in markets that are governed by legal systems that enforce property rights and contracts. These conditions have in mind an advanced capitalist economy rather than the sort of economy and economic environment prevalent in the past. The further back in time we go the less applicable the basic assumptions of neoclassical economics are likely to be. As the anthropologists Mary Douglas and Baron Isherwood argued many years ago: 'the view of the world organized as a competitive power-seeking game between individuals exhibits a cultural and temporal bias'.³⁹ In other words what is often presented in economics as a general rule of behaviour and as the tendency of 'human nature' at all times and places (where large numbers are involved) is in fact only approaching a reality in certain times and for certain cultures. Even in modern advanced capitalist societies, 'Considering that most of us wander in a fog of indecision and emotion the bright sunlight in which the rational man strides toward his goal is difficult to credit'.⁴⁰

Most historians and certainly most anthropologists would argue that economic rationality should be regarded as a variable and not as a fixed postulate.⁴¹ In past societies it may have been more rational for individuals to act to maximize the economic position of the village, the estate or the extended family than the individual or the nuclear family. Alternatively the most 'rational' behaviour in some past societies may have been not to maximize material welfare at all: sufficiency may be more important than maximization. Where mortality is high and disease rife and/or where there are few consumer goods available, leisure rather than wealth might be a prized possession. Thus the problem, as Coleman has argued, is that:

For the historian, rationality in the choice between alternatives is not a necessary assumption of human behaviour; for the economist, it is crucial to the proper functioning of models of that behaviour.⁴²

Marshall Sahlins's study of various hunter-gatherer groups in Africa, Australia and the Pacific demonstrates that these peoples would be regarded as both poor and lazy from the point of view of Western economic science but in reality they were working to a different set of priorities and goals that were more rational to them in their circumstances. They were relatively affluent in regard to their needs and wants with varied diet and

plenty of leisure time. Sahlins takes to task the dominant definition of economics derived from Lionel Robbins in the 1930s: '[economics is] the logic of choice . . . the study of how scarce resources are allocated among competing ends.' Sahlins stresses that ends should not be taken for granted as insatiable or unvarying and that scarcity is only understandable in relation to a society's wants and desires 'it is not that hunters and gatherers have curbed their materialistic "impulses"; they simply never made an institution of them'.⁴³

In addition, market clearing is unlikely to be smooth or efficient and may not even be a tendency in markets of a pre-industrial economy. Even in advanced economies markets with 'sticky' factors of production such as labour fail to clear efficiently or at all. Thus an economic historian needs to be open to alternative ways of conceptualizing and thinking about economic behaviour in the past. The tools and assumptions of neoclassical economics can only be helpful in some circumstances, but in others, and especially in pre-industrial societies, the historian must try to understand the 'cultural otherness' that lies behind motivations and behaviour, with respect to material life, that do not fit with our contemporary experience. Karl Polanyi argued that historians should adopt a substantivist approach to understanding the economies of earlier times. By this he means that they should, like an anthropologist, try to understand past societies and economies in their own terms and not through the lens of present time and culture. He suggests that there are many types of economic action other than market interaction in earlier societies, all of which need to be understood and theorized as much as we have theorized around market behaviour. He cites the importance of householding in the past: exchange and the husbanding of resources at the level of the household. He also emphasizes reciprocity which should be seen as much more than just exchange without money. Reciprocity is driven by different motivations and relationships than those of the market and it certainly does not disappear with the growth of markets as the size and importance of gift-giving sectors in advanced economies testifies.⁴⁴ The redistribution of wealth and well-being through both charities and state-organized taxation is another undertheorized area.

Finally, the dominant paradigm in economics – neoclassical optimization – is not only questionable as a driving motivational force, in most periods and circumstances, but is also not a source of deep insight into discontinuity or change over time. It has always been more concerned with function, balance and equilibrium:

For the economist time is a troublesome intruder, bringing disorder to the symmetry of theory, threatening the exactitude of the short term, conveniently defined by the *ceteris paribus* assumption that other things indeed will remain equal. The historian's interest however is almost always with the long run when things will not remain equal but will change in indefinable ways. So the precision of the models is secured by the drastic device of omitting any attempt to evaluate the effect upon economic change of non-economic influences, be they technological or political, cultural or demographic: a very high price to pay for a particular achievement.⁴⁵

Econometric tools and econometric history today

The most important quantitative history in Britain immediately following the first wave of econometrics avoided theory to a large extent in favour of counting and estimating: national income accounting and estimations of the growth of output, national income and population change became an important preoccupation (for more detail on this see pp. 257–259).⁴⁶ There has been a continuing important thread of econometrics in the economic history of Britain but nothing quite as ambitious as in the first wave. American economic history by contrast has remained more committed to econometric approaches and much progress, both theoretical and technical, has been made.⁴⁷ It is the case that contemporary economic theory is pushing out the boundaries of what can and should be incorporated in modelling the past. The development of economic theory has resulted in the ability to create models that are much closer to simulating the variety and complexity of real world situations and that are much less hedged in by stylized facts than in the past. The technical limits of economics have been greatly extended. The advances can be seen particularly in new classical macroeconomics; new (endogenous) growth theory; new industrial economics; new international economics; multiple equilibria theorizing; new theories about invention, non-stationary time series. The greater sophistication of rationality assumptions and the ability to allow for changing tastes and conventions might also be stressed whilst developments in economic geography and in economic sociology are helping to create economic models that can be applied to a larger variety of circumstances.⁴⁸ All of these developments have served further to highlight the crude nature of the first wave of econometric history and to suggest that the future for cliometrics might be much brighter.

Economic history itself has had a role in undermining neoclassical growth economics by emphasizing the importance of factors internal to economies in furthering or hindering the growth process such as ‘social capability’, the quality of ‘human capital’, market size, institutions, tax regimes and capital accumulation strategies.⁴⁹ Not all cultures or countries react in the same way to international market forces or the global availability of new technologies: catch up and convergence are not automatic; divergence and overtaking are more likely. Thus endogenous growth models have developed that are more attuned to the circumstances and nuances of different parts of the globe, experiencing different economic and historical circumstances. Another area where quantitative economic history is working alongside growth economics is in developing more sophisticated measures of living standards that incorporate life expectancy, infant mortality, literacy levels, political and civil rights, as well as real income per head. Although there are problems with quantifying and weighting such components, the Human Development Index (HDI) is likely to contribute much to comparative economic history in the twenty-first century.

Economic theory today is seldom based upon models of perfect competition. Neither is supply and demand analysis based upon unquestioned marginal equivalences or on unproblematic information flows. The neoclassical stress upon the importance and computability of shifts in supply, demand or price at the margin has been replaced in

historical work in particular by acknowledgement of customary and inflexible elements in wages and prices. More attention is also now paid to **asymmetric information** where one party in a transaction knows more than the other.⁵⁰ This asymmetry is likely to have been much more important in the past especially in distant trading. It also occurred between lenders and borrowers in regions and cities in the eighteenth century when it was mitigated by the activities of notaries and attorneys. In an example from more recent history, Martha Olney's article on consumer credit in black and white households in US cities (see exercise on p. 239) uses the theory of asymmetric information to sophisticate her supply and demand model. Lenders faced 'moral hazard' risks as debtors could always behave imprudently and default. The pool of applicants for credit may also have had proportionately more bad credit risks than the population as a whole, so lenders additionally faced 'adverse selection'. Creditors thus screened applicants partly on the basis of race but were less worried regarding instalment credit than shop credit because risk was reduced by the ability to repossess consumer durables.⁵¹

Reinforcing theories of asymmetric information is acknowledgement of agency problems. **Agency theory** is concerned with how people get others to do what they want where there is asymmetric information and where incentives and monitoring will be required. Insight from the economic literature on efficient contracts enabled Ann Carlos to reconsider the level of opportunism among Royal Africa Company agents in the seventeenth century and to research, in a similar way, the agency problems of the Hudson's Bay Company in a joint article with Steven Nicholas.⁵² The related concepts of adverse selection and moral hazard in decision-making by economic actors go some way towards creating a more realistic set of assumptions for the application of economic theory to the past as well as to the present. Insights from this area of theory are important in providing tools for analysis of the growth and operation of long-distance trade and of relationships within and between firms.⁵³

Cliometricians are now also much more concerned than in the past with considering how people bargain and cooperate on the ground; the role of institutions; and how transactions costs shift to impact upon human economic interaction.⁵⁴ The importance and overlap of economic, social, familial and communal networks have been stressed by economic sociologists, geographers and economists alike and the importance of local cultures and understandings in underpinning the dynamism of localities within a wider framework of global exchange is now firmly recognized.⁵⁵ Gender perspectives are also coming into play to a greater degree than in the past and have the potential to undermine a raft of traditional assumptions underlying economic theory. In particular the application of standard neoclassical tools to the so-called 'new household economics' of Gary Becker and his followers has come under attack for ignoring intra-household bargaining games and for endorsing existing household structures as 'efficient'. As Jane Humphries has argued: 'despite the changes neoclassical economics has been undergoing, so long as it continues to privilege the individual over the social in the hierarchy of causation, then it must assume that whatever exists must be optimal, otherwise it would already have been changed'.⁵⁶

Rational choice has been the dominant paradigm in neoclassical theory and has strongly influenced other social sciences but empirical investigation more recently has

History by Numbers

increasingly shown that choice is biased and that rationality is bounded.⁵⁷ Choice is naturally myopic: biased towards the present and against the future. It is oriented towards the aversion of risk or loss and in favour of fairness. This implies that preferences will often be reversed over time and that choice is likely to be time inconsistent. The concept of time inconsistency is most often used where a rational private sector is aware that a policy maker has an incentive to renege on a policy and takes this discovery into account by changing behaviour.⁵⁸ But, as Kenneth Arrow has argued:

The very concept of rationality becomes threatened ... [when] perceptions of others, and in particular of their rationality become part of one's own rationality. Even if there is a consistent meaning it will involve computational and informational demands totally at variance with the traditional economic theorist's view of the decentralised economy.⁵⁹

Rational action is affected by many circumstances: by power relationships as well as by the degree of information to which economic agents have access (hence their expectations) and by their cognitive ability in processing such information. It is also influenced by an array of more complex psychological factors that are essentially 'irrational' and remain ill-researched in this context.⁶⁰ The assumption of market equilibria and market behaviour will always be prone to challenge, particularly in societies with poor information flow, devolved power structures, forceful customary arrangements and low levels of literacy and education. For these reasons, as a recent survey points out 'Trust, uncertainty, creativity, credibility, institutions, agency and informational asymmetry are all concepts that have been assimilated into the cliometricians' toolbox'.⁶¹

Since the 1980s economic theory has become enormously more sophisticated as has the power of computer-aided multivariate modelling. Econometrics has also moved much closer to statistical modelling and statistical method and is less tied to the use of mathematical models. Mathematical economics was a prerequisite for certain applications of statistics in economics, particularly in cross-sectional, market equilibrium modelling, but there is no necessity for the two approaches to go together. Increasingly they have not and the cutting edge of econometrics and econometric history today, for example game theory, relies much more on statistical than on mathematical methods. A grasp of price theory and its algebraic modelling is no longer the defining skill of the cliometrician. Increasingly sophisticated techniques of econometric and statistical analysis are to some extent providing a new springboard for the application of economics to history whilst developments in economic history, especially in time series and panel data analysis are having an influence upon the economic analysis of contemporary evidence.⁶²

In summary, future relationships between economics and history look potentially more fruitful than at any time since the 1970s. This is partly because the preoccupations of some economists are changing and they are more in line with the central concerns that economic historians have had for many years. Such concerns as the varieties of market and other forms of exchange behaviour and practice; 'irrational' behaviour, the variation of performance of firms under similar sets of conditions; the structure of trade,

the role of trust and reputation, economic horizons in risk taking and decision-making, backward and forward linkages, economic and financial crises, competitiveness, migration, living standards, nutritional levels, welfare reform. The new potential for relationships between economic theory and history is also boosted by the development of more varied and more sophisticated techniques of computer-aided statistical analysis.⁶³

Comparative economic growth and new time series research

As the first wave of New Economic History died away, the cutting edge of quantitative economic history started to be associated with the measurement and analysis of economic growth. Starting with the production of new estimates of output and growth trajectories of advanced nations and later turning to the statistical analysis of the causes of change in different global contexts, this topic has remained at the forefront of quantitative economic history ever since and requires separate attention from us here. It is important because it has provided new opportunities for economic historians to engage with and to test the models and ideas of growth theorists and of theorists of innovation and technological change.

As early as the 1960s, but gathering pace in the 1980s and 1990s, growth rates for the British economy from the eighteenth century to the present were undergoing revisions.⁶⁴ Thanks to the new output figures based upon weighted sectoral index numbers, the classic Industrial Revolution began to appear much more gradual than was previously thought. Although there was much debate about the data extrapolations, proxy figures and the ‘index number problem’, that is, the weights of the sectors (that are largely based on the distribution of male occupations drawn from parish register figures), there is no doubt that our ideas about the overall pace of industrialization were permanently revised by this work.⁶⁵

Analysis of the growth trajectory in Britain, and elsewhere, initially took the form of preoccupation with explaining the ‘residual’ exposed by input compared with output measures; a residual likely to indicate shifts in productivity caused by technological and other innovations.⁶⁶ But new endogenous growth models were developed in the 1980s and 1990s that attempted to explain growth without placing primary emphasis on exogenous variables such as technology.⁶⁷ Human capital formation, the private rewards accruing to innovation, the relationship between private and public knowledge have since become the leading variables in historical as well as contemporary analyses of comparative economic growth.⁶⁸ In addition, there has been growing concern to theorize and to model the role of geography and natural resources in development, including forms of path dependency, triggered by resources.⁶⁹ Path dependency has also become associated with institutions that do or do not favour democratic regimes and the operation of free markets.⁷⁰

In earlier chapters we used some examples from this recent international research on global comparisons of economic growth and further articles relating to comparative economic growth and the likely causes of differential growth are included in the exercises

History by Numbers

relating to this chapter. Much of the effort of such research has been focused upon explaining the so-called Great Divergence in growth rates between the West and Asia since the eighteenth century. Why was Western Europe, and Britain in particular, the location of the first major transition to modern (technologically innovative and sustained) economic growth? Renewed debate on the precocity of the West, compared with China in particular, was sparked by the publication of Kenneth Pomeranz's book *The Great Divergence. China, Europe and the Making of the Modern World Economy* in 2000 where he argued that the level of development in China was very similar to the West until the eighteenth century.⁷¹ The assertion was quickly disputed and an array of new growth estimates and discussions regarding Asia and the West has been the result.⁷²

Such research centres mainly upon national income accounting that sums the weighted sectoral outputs of an economy (using composite indices), in relation to population totals, in order to measure and to compare GDP per capita. The recent comparative research of Broadberry et al. (featured on pp. 138–139) is much in this mould. The major source of figures employed for this and other recent comparative growth research is the continuously updated GDP per capital figures collected in the Groningen Database at the Groningen Growth and Development Centre.⁷³ Their specialization in this respect has recently developed outwards (in the CLIO-infra project) from GDP to a wide range of other historical developmental indicators covering biological, institutional, environmental and human capital indices.

Most recently theories of directed technological change have become current. These place primary emphasis not upon culture or institutions but upon technological change being sparked by differential factor costs occurring in particular circumstances. Transnational comparisons of factor costs have been highlighted, particularly the importance of levels of real wages in determining the degree of substitution of capital for labour (embodied in new technology) in the production process. Leaning heavily upon the concept of directed technological change originally promoted by Daron Acemoglu, Robert Allen has suggested that the primary cause of divergence in rates of economic growth globally in the modern period was the real cost of labour relative to the costs of capital and fossil fuels. In high wage economies, such as Britain in the eighteenth century, inventors and innovators concentrated upon reducing the need for manual labour whereas similar innovation was held back by the cheapness of labour in other parts of the globe, particularly in Asia.⁷⁴

When considering these sorts of research prompted by comparative panel or matrix data expressing composites of composites, and elaborately contrived 'real' indices, readers should not only be vigilant about the level of accuracy of the evidence for each country/region but should also ask whether figures are directly comparable. Lack of commensurability of data across time and space often compromise comparative figures. Data is always drawn from the recorded economy which means that conventional yardsticks of growth are most seriously distorting for economies with a large informal or (inadequately recorded) self-employment sector.⁷⁵ Information is likely to be fuller for economies where the state is strong and keeps good records for fiscal and other purposes. This leads to the problem of circularity in that measures such as GDP and real wages

are sometimes a better indicator of the strength of the state than of real economic change.⁷⁶

Finally in appreciating the potential pitfalls as well as the usefulness of such research it is important for the reader to consider the way in which a common deflator or denominator has been used to compare the values placed on output or real wage data across time and space. For GDP the most commonly used denominator is the 1990 Geary Kamis dollar (GK\$). In 1990 the UN calculated the poverty line at a dollar a day, so comparing each country's output with the purchasing power that it represents (per capita) in 1990 dollars makes it very easy to see the differences across time and space. It has thus become a convention even though there is little justification for sticking with 1990 dollars when other comparators generate rather different results.⁷⁷ For real wage measures over great distances and long time periods a silver standard is generally used but this is also very questionable as the market for silver varied and was poorly integrated within and between regions.⁷⁸

The models, the evidence, the reality

At the heart of general methodological criticisms in economics and hence in econometric history has been the problem of *instrumentalism*, in other words, judging theories on the basis of their predictive ability. As Crafts has pointed out, for example:

Given that economists are often concerned with prediction rather than explanation up to a point this may be an acceptable criterion and, for example, firms may be regarded *as if* they maximise profits if their actions are not inconsistent with profit maximisation. For an economic historian this may be a dangerous oversimplification leading to an erroneous belief that motives have been understood or that all decisions are based simply on profit maximisation.⁷⁹

A further methodological problem is that the hypotheses derived from neoclassical theory are rarely subjected to tests that can falsify them.⁸⁰ Because there are so many stylized facts and *ceteris paribus* elements, failure of a model to line up with the evidence is always easy to explain away in terms of the need to adjust the parameters of the model in some way. Prior subjective belief in models is almost always present. Chicago School economists are, for example, readier to accept empirical results consistent with standard price theory and less willing to believe in market failure than those trained elsewhere.⁸¹ In addition most historians believe that historical statements of causation are always necessarily weaker and more complex than those of economists and that they are unsuitable for testing by the economists' method of isolating just one cause and assessing the outcome *ceteris paribus*. The concept of causation and the closed and deterministic nature of economic models are therefore both regarded with extreme scepticism. As Sir John Habakkuk put it long ago: 'one man cannot think in two ways'. The economist's search for clearly specified models of collective behaviour, susceptible to test by

History by Numbers

quantitative methods, leads him to approach history in a way which differs fundamentally from the search for the sources of individual behaviour that is characteristic of the work of many historians.⁸²

One of the main difficulties with all forms of quantitative history is that the original data may be too unreliable, biased or incomplete to allow any meaningful manipulation of the figures. When the statistical testing of economic or other models is the central point of the research such data problems need to be particularly carefully considered. Any unreliability or bias may become magnified if variables in a model are relied upon heavily for predictive or extrapolative purposes. In the enthusiasm for modelling it is especially easy to give too little consideration to how the original evidence was gathered and categorized, what elements may be missing from it and what other forms of distortion it may contain. The problems are compounded where, as is often the case, proxy variables are used because the information which the historians would ideally like to have is not available. Sometimes proxies are sought which may have too tangential a relationship to the variables under consideration:

Lacking output data we use trade figures which are themselves based upon customs returns or other tax statistics; parish register and hearth taxes stand for vital statistics; excise figures for industrial output. The figures once gathered in, invite processing. So they are put into time series, tested, correlated, made to yield growth rates, or coefficients of one sort or another, find their way into equations and end up in computer programs. Some assumptions of economic rationality are made; and out come the results to two places of decimals. Scientific history gets a boost; positivism is reborn. The cautious stress that it is all probabilistic, a mere step towards truth via the formulation and testing of models.⁸³

Conclusion

We have seen that many advantages can be derived from applying quantification in history but that additional vigilance is required in marrying quantification with the construction and testing of models and concepts based on economic theory. We have considered the nature of neoclassical analysis and the potential problems in applying neoclassical assumptions to historical contexts. Newer developments in theory have been outlined that are creating a rejuvenation of the econometric cause. Whether this will flower into a new wave of historical econometrics is debatable although the growing sophistication of computer-aided statistical applications is in place to encourage such a trend. The real challenge for econometric history is how to incorporate new variables into meaningful formal computational models with the attendant problem of deciding in which circumstances it is no longer useful to do so given the complexity of the model and the problems of source materials. It is interesting that most of the historical work incorporating efficient contracts theory, asymmetric information, institutional structure, gender and other factors has relied very little on the statistical manipulation or testing of

models. In order to create a new momentum in econometric history it will be important for historians to avoid the mistakes of the past: carelessness with sources and with questioning the data, preoccupation with macro-level theorizing, anachronistic application of present centred theory to the past, the free-market, optimization bias, studies of men not women, nations not regions or localities. A further problem attends the currently popularity of studies of comparative economic growth. We have already noted that estimates of GDP per capita are subject to errors and bias that become particularly important in (national) comparative work because one is rarely comparing like with like. Additionally one might also ask whether GDP per capita is the most appropriate yardstick against which to judge the economic success or performance of economies. GDP per capita takes no account of the distribution of income, of inequality, poverty or other health or social indicators of development and it has no concern for sustainability. New work using a wider array of indicators of development, associated with measures such as the Human Development Index (HDI) or proposing an entire rethink about the trajectory of development and its impact upon inequality point a way forward.⁸⁴ Although these approaches are in turn subject to debate and criticism regarding the commensurability of transnational data and the development theories that they employ, they open the way to increasing use of metrics and quantitative methods in economic history in future years.

Further reading

- Coyle, Diane, *GDP: A Brief But Affectionate History* (Princeton 2014).
- Greasley, David and Les Oxley (eds), *Journal of Economic Surveys*, 24 (5), (2010), Special Issue on Econometric History.
- Jerven, Morten, *Poor Numbers: How we are Misled by African Development Statistics and What to Do About It* (Ithaca 2013).
- Klein, Judy L., *Statistical Visions in Time: A History of Time Series Analysis 1662–1938* (Cambridge 1997).
- Lamoreaux, Naomi, ‘Economic history and the cliometric revolution’, in A. Molho and G. S. Wood (eds), *Imagined Histories: American Historians Interpret the Past* (Princeton 1998), pp. 59–84.
- Lamoreaux, Naomi, ‘Beyond the old and the new economic history’, in Francesco Boldizzoni and Pat Hudson (eds), *The Routledge Handbook of Global Economic History* (London 2016) pp. 35–54.
- Lyons, John S., Louis P. Cain and Samuel H. Williamson (eds), *Reflections on the Cliometrics Revolution. Conversations with Economic Historians* (London 2008).
- McCloskey, D. N., *Econometric History* (London 1987).
- Morgan, Mary, *The History of Econometric Ideas* (Cambridge 1990).
- Philipsen, Dirk, *The Little Big Number: How GDP Came To Rule The World and What To Do About It* (Princeton 2015).
- Ward, Michael, *Quantifying the World: UN Ideas and Statistics* (Bloomington, Indiana University Press 2004).

CHAPTER 9

HISTORICAL RESEARCH, COMPUTING AND THE DIGITAL REVOLUTION

The rapid growth and current mass of quantitative and qualitative historical data available online is transforming the sources of evidence available to historians and has fundamentally altered research possibilities and practice in the last few decades. It is now possible in most fields of history to research and publish using online sources exclusively. Historians can pursue careers without close contact with the nuances and difficulties of dealing with primary sources directly, and divorced from the taxonomies of knowledge built up over earlier centuries.¹ At the same time computer hardware and software has become infinitely more sophisticated and more widely available than in the early days of the application of computing to history. The aim of this chapter is to consider the impact of these two connected developments upon historical research and writing and to provide an introduction to broad types of computer-aided analysis in historical research. We consider this as an aspect of 'history by numbers' and look at the pitfalls as well as the advantages that have arisen from the major extension of computer use in history in recent decades.² We focus upon the difference that the employment of digital sources and computer technology makes to historical research practice and results. There is also a practical section that one can turn to immediately for advice with using spreadsheets, databases or other software in historical project work. This should assist researchers in getting started with a historical research project and will advise on where to look for more detailed, specialist help and for further illustrative examples.

First a brief survey of software types is useful especially for those encountering digital sources and computer applications for the first time in relation to historical research.

Useful software types

The types of software used by historians in the storage and retrieval of information and in quantitative research fall into six broad types though they generally overlap, especially with the application of interfaces that enable the movement of data from one to another:³

- (a) Spreadsheets
- (b) Statistics and graphics packages
- (c) Databases and database management systems
- (d) Textual processing packages
- (e) Software for qualitative data analysis
- (f) Spatial and mapping software: GIS (Geographical Information Systems).

History by Numbers

We will consider each, briefly, in turn.

(a) Spreadsheets

A **spreadsheet** is a computer application for the processing and display of statistical information. It allows the storage of numerical and textual data in matrix format and enables a variety of statistical and graphical manipulations to be performed. Although they were first developed for business and commercial applications, the use of spreadsheets in academic studies has been growing rapidly in recent decades. Historians, amongst others, were quick to see the utility and the potential of this highly flexible class of computer applications.⁴ In historical research they combine some of the functions of statistics packages and of database software (for both see below) but are normally much simpler than either of these to use even if they are ultimately more limited. For most storage and retrieval of information and for most analytical/statistical tasks that historians might need, a spreadsheet is likely to be the tool of choice. Excel is perhaps the most popular spreadsheet software currently as it is part of the 'monopoly' created by Microsoft Office. There are many other examples, some open source (freely available via the Internet) and others available as part of larger software packages geared to different purposes.

Spreadsheets enable information to be tabulated under a variety of headings and subheadings and they are particularly useful for large datasets, to show and examine change in data over time and for comparative study of different sets of information. Much of the data collected by governmental agencies including those sources openly available online, via the Office for National Statistics, for example, are stored in spreadsheet format and can be downloaded already in such formats for further analysis and manipulation.⁵ In elementary statistical processing, spreadsheets have a big advantage of speed and accuracy over manual techniques. They can be used easily to produce measures of central tendency and dispersion, frequency distributions, scatter graphs, bar charts, pie charts and line graphs. Many spreadsheets incorporate statistics add-ons and can perform simple correlation and regression analysis and operate random sampling techniques and tests. If more sophisticated graphs or figures are required it is usually possible to use a graphical user interface or GUI. The main problem with spreadsheets is that virtually all were developed for business or home accounts, not for historical use, the default settings are often inappropriate for historical work and adjustments are often necessary. That said, for most quantitative historical research, a spreadsheet will store the required information in tabular form and enable one to use practically all of the quantitative methods covered in this volume. Spreadsheets are accessible and easy to use for historians with very limited IT or statistical training.

(b) Statistics and graphics packages

Statistical and graphics packages designed for general or social science applications have been widely used in history for several decades. Dominant in many quarters from the late 1970s to the mid-1990s was SPSS (Statistical Package for the Social Sciences) initially with its variants SPSSx, for more powerful machines, and SPSS-PC for personal computers. The big attraction was that it was geared to social science applications and to

projects that collected similar data and asked similar questions to those posed in much historical research. The disadvantage was that a lot of coding of information was required prior to processing. The same problem applied to Statistical Analysis System (SAS), a similar package. For most simple statistical exercises smaller and less complicated software was often preferred even though it did not have a specific social science purpose. One such was Minitab, developed at Pennsylvania State University, popular in the 1980s and early 1990s. This package was used by students studying the course upon which the first edition of this book was based and it is still currently available (little changed in essentials) in version 17. There are however many good and much more recent statistical and econometric packages available for research at many different levels of technical sophistication, such as EViews and Stata.

Although both SPSS (owned and developed by IBM since 2009) and SAS remain in use in advanced analytical statistics in large data projects such as in the health sciences, in the last two or three decades most history researchers have turned away from dedicated statistical packages, for social science or academic use, regarding them as generally too wide-ranging and complicated for the relatively simply descriptive and summary statistics in which most historians are engaged.

(c) Databases and database management systems

A **database** is a collection of related data, organized in a predetermined manner and according to a set of logical rules. Such data can be stored in a computer in machine-readable form. A **database management system** (DBMS) is a computer programme that allows one to organize and manage the information stored in the database. A DBMS is used for storage, sorting, retrieving and interrogating numerical and other information that is inputted and arranged in a series of tables in the database. As with spreadsheets the data are arranged in electronic tables but databases differ from spreadsheets in performing complex tasks of selective reordering and retrieval of information (like a sophisticated filing system) and in being able to handle, process and retrieve non-numeric, as well as numeric, data in the cells of the matrix more easily than a spreadsheet. Currently available database programs can handle several pages of text in each record and can use these texts in searches. They can now also include images and live links to documents and files outside the DBMS. Examples of these are Access, Paradox, Foxbase and Dbase 3. These have been continuously updated over the years, are easier to use than in the past, can be customized for views, browsing and reporting and are more ‘intelligent’ in detecting inconsistencies, defective structures and illogical queries than their predecessors. The new wave of DBMS also facilitates movement of data and results across systems, to statistical analysis and graphical packages, and to spreadsheets.

Each row of a database table is a unique record relating to a case. Each column contains one ‘field’ of information relating to the cases. This is illustrated in Figure 9.1, drawn from a research project comparing two Yorkshire townships in the long eighteenth century. It is a screen from the inputted land tax return for Calverley in 1784 which can be compared with a folio from the original source in Figure 9.2. Using a DBMS enables selective information about all cases or just some particular cases to be called up and

History by Numbers

The screenshot shows a Microsoft Access database window with a table titled 'Calverley - Table'. The table has columns: refno, year, pfname, sname, ptstatus, osname, ostatus, pounds, shillings, and pence. The data consists of 348 records, mostly from 1784, listing names like Thomhill, Atkinson, Ward, Hawsorth, etc., with their respective status and financial details.

refno	year	pfname	sname	ptstatus	osname	ostatus	pounds	shillings	pence
324	1784	Thomhill	Thos	Esgt	Atkinson	Joseph	1	6	4
325	1784	Thomhill	Thos	Esgt	Carter	Joseph	0	6	8
326	1784	Thomhill	Thos	Esgt	Ward	Samuel	0	17	8
327	1784	Thomhill	Thos	Esgt	Hawsorth	George	0	10	8
328	1784	Thomhill	Thos	Esgt	Ires	Thomas	0	9	8
329	1784	Thomhill	Thos	Esgt	Kaighley	John	1	4	0
330	1784	Thomhill	Thos	Esgt	Carter	George	1	12	4
331	1784	Thomhill	Thos	Esgt	Ross	Benjamin	1	5	9
332	1784	Thomhill	Thos	Esgt	Wilson	John	0	9	8
333	1784	Thomhill	Thos	Esgt	Williamson	John	0	7	0
334	1784	Thomhill	Thos	Esgt	Turner	Robert	0	6	0
335	1784	Fisher	Jonathan		Turner	Robert	0	3	10
336	1784	Atkinson	Joseph		Turner	Robert	0	0	10
337	1784	Atkinson	Joseph		Huttler	Joseph	0	0	0
338	1784	Atkinson	Joseph		Thrapleton	James	0	6	8
339	1784	Mather	Joseph		Roberts	John	0	2	8
340	1784	Greenwood	William		Hollings	Thomas	0	4	4
341	1784	Carter	George		Smith	James	0	6	0
342	1784	Baker	William		Baker	William	1	1	2
343	1784	Marshall	Abraham		Marshall	Abraham	0	1	2
344	1784	Overend	Jonas		Clarkson	William	0	10	0
345	1784	Overend	Jonas	Mr	Child	James	0	2	8
346	1784	Wormit			Lister	Joseph	0	8	8
347	1784	Hawsorth	Abm		Hawsorth		0	1	6
348	1784	Nicholls	William		Nicholls	William	1	6	8

Figure 9.1 Screen from Calverley land tax return, 1784.

Source: Pat Hudson and S. A. King, research project on two textile townships.

analysed. Thus the researcher is able to create further secondary tables or units of analysis geared to the research question being posed, usually by breaking down the information and joining them in an order different from that occurring in the original sources.

Databases can be **flatfile** (simple, and usually composed of just one table) or **relational** (involving many tables with the possibility of linking data across files and interrogating several tables simultaneously). The land tax screen is one of many tables from a larger relational database that enables information about individuals to be drawn from many different tables/sources. Thus the land tax payers and occupiers can be linked to other records such as militia lists or parish registers. Table 9.1 lists the different major files contained in a relational database relating to the Yorkshire townships project. The aim was to use nominal record linkage to work outwards from family reconstitution to a more detailed reconstruction of the circumstances and life courses of ordinary people. A similar early project of reconstruction is provided by the Labour Markets Database at Queen Mary and Westfield College. This is not concerned with nominal linkage but with mapping the geography of economic distress before 1914. The use of a relational database here makes it possible to use a variety of indicators of distress for each geographical region. The database has recently been expanded to include more twentieth-century data and health and welfare indicators.⁶

Name of proprietor	Name of occupier	Land	House	Other
Mr Thornhill Esq.	Joseph Atherton	3	6	4
D.	Joseph Carter		6	8
D.	Samuel Ward	17	8	
D.	George Flannworth	10	8	
D.	Thomas Hob	9	8	
D.	John Heightley	1	4	2
D.	George Carter	1	12	4
D.	Benjamin Mofe	1	5	9
D.	John Wilson	9	8	
D.	John Williamson	7	8	
D.	Robert Turner	6	0	
John Fisher	Robert Turner	3	10	
Joseph Atherton	Robert Turner	-	10	
D.	Joseph Fletcher			
D.	James Thrapleton	6	8	
Joseph Mather	John Roberts	2	8	
William Greenwood	Thomas Hollings	4	4	
George Carter	James Smith	6	0	
William Baker	Hinself	1	1	2
Abraham Marshall	Hinself	-	1	2

Figure 9.2 Page from the land tax return for Calverley with Farsley, West Yorkshire, 1784.
Source: West Yorkshire Archive Service, Wakefield.

With a relational database it is very important to specify fields of information, for example, surname, forename, date, place in the same way in all tables and each case in each table must have a unique reference number. Field lengths for the same variables must be the same across all tables and there must be consistency in the use of characters or integers (letters or numbers). This is because the computer is only able to make links by comparing like with like. Where spellings vary, for example with surnames or where there may be many descriptions for the same occupation, it will generally be necessary to add fields that group or code such variations (for example, in surnames or occupations). The fuzzy searching required to make a link is often built in to the software. Table 9.2 gives a design for baptism entries for the seventeenth to the nineteenth centuries whilst

History by Numbers

Table 9.1 Details of a selection of computer files on Sowerby township in a relational database

Document	Filename	Fields (no.)	Entries (no.)
Poor rate assets, assessments 1738–1855	ARRP	10	5 539
Parish register baptisms, 1668–1825	BAPTISMS	19	13 874
Bonds of indemnity, 1755–1781	BONDS2	24	27
Parish register burials, 1699–1837	BURIALS	23	11 211
Rate Assessment, Blackwood, 1804	BWOODRA	14	1 178
Hearth Tax 1664, 1666, 1672, 1674	HEARTHTAX	8	1 822
Independent baptism register, 1740–1837	INDBAPS	19	1 205
Land Tax assessments, 1750–81	LANDTAX1	11	1 392
Land Tax assessments, 1782–98	LANDTAX2	14	3 540
Land Tax assessments, 1799–1800	LANDTAX3	17	612
Parish register marriages	MARRIAGE	17	90
Methodist baptism register, 1790–1837	MBAPS	18	337
Militia Tax, 1716	MILITIA	11	287
Settlement exams, 1744–1808	SETEXAMS,	15	52
	NEWEXAMS	16	52
Pauper Apprentice records, 1720–1801	PAUPERAPS	17	551
Household Census, 1764	POPBOOK1	17	597
List of proprietors and occupiers, 1827	PROPLIST	13	335
Removals, 1712–1751	REMOVALS	9	5
1811 Census	SCENSUS	11	461
Settlement certificates, 1688–1749	SETTLEMENT	13	170
Apprentice indentures, 1721–1801	TRADAPS,	20	78
	TRADEAPS	21	80
Window Tax, 1759	WINDOWTAX	8	135
Probate inventories, 1689–1785	SPROBATE	14	111
Wills, 1689, 1785	SWILLS	20	111

Source: Pat Hudson and S. A. King, research project on two textile townships.

Figure 9.3 shows a page from the corresponding primary source, the parish register. The number of fields and field lengths are designed to cover all cases. There was, for example no grandparent details in the source until the end of the eighteenth century and few residence or forename fields actually needed more than 12 characters, 30 was assigned as a comfortable maximum. Fields have been added to code surnames and occupations. In this case two fields have been added, where available, from other sources to assist in identifying correct links. The T-score fields indicate the probability that a correct link has been made between the baptism entry and the marriage entries for both the parents and for the baptized child. Linking baptisms in this way is part of the process of computer-aided family reconstitution widely used in historical demography.⁷

The most common interrogation format used with a DBMS is **ISQL** (Interactive Structured Query Language) which enables the calling up of certain vectors of information

Table 9.2 Design for a family reconstitution database file of eighteenth-century baptisms

Fields	Function	Length ^a
Refno	Unique identification number	6
Bd	Birth day	2
Bm	Birth month	2
Byear	Birth year	4
Sx	Sex	1
Fname	Forename	15
Sname	Surname	20
Status I	Birth date	10
Status II	Status	10
Residence	Residence	30
Ffname	Father's forename	15
Fsname	Father's surname	20
Focc	Father's occupation	20
Fstatus	Father's status	10
Fres	Father's residence	30
Mfname	Mother's forename	15
Msname	Mother's surname	20
Mocc	Mother's occupation	20
Mstatus	Mother's status	10
Gppat	Paternal grandparent	35
Gpmat	Maternal grandparent	35
Psnamecode	Surname code of child	4
Fsnamecode	Surname code of father	4
Msnamecode	Surname code of mother	4
Tscoremdb	Refno of parent's marriage	8
Tscoremds	Refno of own marriage(s)	8
Tscormdg	Refno of kinship-linkedmarriages	20
Wbfamily	Social status indicator	1
Mfamily	Landholding indicator	1
Tscore	Confidence flag	2

^a Number of characters/integers allowed.

Source: Pat Hudson and S. A. King, research project on two textile townships.

relating to particular cases in the table or tables, the ordering and summing of such vectors and other more complicated commands involving highly selective cells of information and the linking of one record with another, often across different datasets within the database. Where a relational database is employed, dedicated software can be added which specifies the linkage criteria to be used in various searches or analysis. These structures of linkage criteria are called **algorithms**. Many algorithms have been written for individual historical research projects and some form the basis of software packages which have been designed to be sold or disseminated for specific historical applications. The Family

History by Numbers

		Baptized 1734
Janu.	10	Isaac son of John Beever of Sowerby West tower
13	11	John son of Mr. Wm of Sowerby Weaver
19	12	George & younger brother of Richard & wife of Blackwood Weaver
24	13	John son of Mr. Wm of Sowerby & par officious man of Blackwood
26	14	John son of Mr. Wm of Sowerby Weaver
Feb.	6	Mark son of Matthew & Margaret son of George & his wife
8	7	George son of George & his wife
12	8	John son of James Meldow of Sowerby Weaver
18	9	Sarah daughter of Henry Claydon of Sowerby Weaver
23	10	Sarah daughter of John Chetham of Blackwood Weaver
Mar.	8	John son of James Heape of Blackwood Weaver
		Baptized 1735
Apr.	6	Mary daughter of Joseph Hollisworth of Westfield Weaver
7	7	Sarah daughter of John Smith of Blackwood Weaver
19	8	James son of William Allen of Westfield Weaver
20	9	Joseph son of Daniel Hollis - of Westfield Weaver
May	8	Matthew son of Richard Newkirk son of John of Westfield
12	9	Grace daughter of Joseph Hill of Westfield Weaver
12	10	Elizabeth daughter of Edward Gaskins of Westfield Weaver
16	11	Mary daughter of Charles Barret of Sowerby Weaver
18	12	Elizabeth daughter of Thomas Hale of Sowerby Weaver
19	13	Daniel son of Benjamin Hale of Sowerby Weaver
June	11	Elizabeth daughter of Daniel Allen son of Blackwood Weaver
15	12	Sarah daughter of John Miller of Sowerby Weaver
18	13	Elizabeth daughter of John Miller of Sowerby Weaver
22	14	George son of Benjamin Hollis of Westfield Weaver
July	9	Sarah daughter of Stephen Green of Sowerby Weaver
25	10	William son of Joseph Hill of Westfield Weaver
25	11	Elizabeth daughter of Abram Farmer of Westfield Yeoman
Aug.	19	John son of Mr. West of Sowerby Weaver
17	20	Mary daughter of Henry Shaddeven of Sowerby Weaver
23	21	Benjamin son of Luke of Westfield Weaver
Sept.	7	Joseph son of Samuel Hollis of Blackwood Weaver
14	8	Alice daughter of Richard Farmer of Sowerby Weaver
Oct.	5	William son of John Hale of Sowerby Weaver
12	6	Ann daughter of John Hale of Westfield Weaver
18	7	Braham son of John Hale of Sowerby Weaver
23	8	Ann daughter of John Hale of Westfield Weaver
25	9	Grace daughter of Nathaniel Taylor of Blackwood Weaver
26	10	Love daughter of Thomas Taylor of Blackwood Weaver
Nov.	1	Anna daughter of James Hale of Blackwood Weaver
3	2	Mr. Farm son of William Hollis of Blackwood Weaver
9	4	James son of Mr. Stoyl of Westfield Weaver
9	5	Sarah daughter of Nicolas Hale of Sowerby Weaver
9	6	Mary daughter of Joseph Shaddeven of Westfield Weaver
14	7	Sara son of John Hale of Sowerby Weaver

Figure 9.3 Page from Sowerby Baptism Register, 1730s.

Source: West Yorkshire Archive Service, Halifax.

Reconstitution software developed by the Cambridge Group is one such example as is the software developed by Mark Overton at the University of Exeter to store and analyse early modern probate inventories.⁸

In recent decades **expert systems** have been used to devise algorithms. Expert systems are advanced softwares that employ artificial intelligence to solve specialized problems. They store rules such as linkage criteria and formulate new rules for linkage from the data as more is inputted. The drawback of expert systems is that they depend upon the existence of stable knowledge systems and that they are governed by identifiable rules. Such characteristics are often lacking in historical data and research.⁹

There has been much debate amongst historians about the need to maintain the integrity of the original source when a database is created. Should information be coded before input (as with different descriptors of the same occupation) and should surname, forename and other spellings be standardized? These are important questions given that databases now and in the future are likely to have a number of users. Historians will probably use databases which have been created (often at some expense) by someone else and primarily for a different and particular purpose. Once coding or standardizing occurs problems inevitably creep in. Information which might be useful or crucial to later users of the database is left out, and coding categories can sometimes force the data into misleading or unhelpful boxes. Standardization of spellings can often eliminate fine differences which make it possible to distinguish between one case and another. The scope for error is legion when decisions about coding or standardization have to be made 'on the hoof' during the process of inputting when unexpected information often fails to fit neatly into pre-given categories. Algorithms that look for similar as well as exact matches modify the problem a little but they by no means solve it. A dominant practice is to create the database with as little coding and standardizing as possible. If it becomes necessary to code or standardize in order to speed processing or create algorithms, this is added (rather than substituted for column fields) at a later stage.¹⁰

The main problems of database use are that complex fields of information in the original documents have to be 'shoehorned' into regular matrix structures and that, although their handling of text has improved, most programs lack sophisticated text searching tools. DBMSs are thus mainly suited to structured sources and to cataloguing in particular. The full featured relational DBMS can also be quite cumbersome but there are some intermediate products such as Filemaker Pro and Lotus Approach that are sufficiently sophisticated for smaller scale projects and a lot easier to use. In addition, the growing sophistication of spreadsheets has meant that they have taken over from database systems proper in carrying out a range of storage, retrieval and analytical tasks for historians. Excel can now carry out most database operations and this trend is likely to continue because most simple database formats, such as Access, use a spreadsheet-style matrix of cells.

(d) Textual processing packages

When one moves from number crunching to 'word crunching' software developments in recent decades have been similarly wide-ranging. The need for historians to deal with

History by Numbers

unstructured data and to use complementary sets of primary and secondary sources in research projects has led to increasing use of a range of software products that span the divide between DBMS and textual analysis software. These include text-oriented DBMSs that allow unstructured data to be entered into fields and then searched using ‘fuzzy matching’ and proximity searches. In this way the user is able to create links between documents.¹¹ At the other end of the spectrum are software packages that facilitate information management but avoid database structures. Instead they employ text searching facilities that are used with indexed and tagged documents. Wordcruncher is one such product. It allows an array of search options using keywords, fuzzy matching, proximity searches and queries of the type: retrieve ‘X’ AND ‘Y’, ‘X’ OR ‘Y’, ‘X’ NOT ‘Y’.¹² The main advantages of the index approach over DBMS text searching are speed of operation and the ability to accommodate much longer texts.

It is perhaps surprising that more fully developed textual analysis software such as TACT (Text-Analysis Computing Tools) and the rapidly advancing WordSmith tools, have not yet had much of an impact in historical research. Both of these can execute a variety of ‘stylometric’ analyses (word counts, word associations, collocations), as well as being capable of comparing many documents at once and identifying the hierarchical structure of the texts. By identifying the habitual use of certain vocabulary and phrases they can also be used to help to identify whether two different manuscripts have been authored by the same person.¹³ The main problem is that historians are not primarily concerned with the use of words or strings of words in a document so much as with the ideas and concepts that lie behind them and which the computer is less able to handle. Such software is also most useful for projects where a distinct body of writings are being analysed rather than the mass of multi-source evidence with which historians most often deal.

(e) Software for qualitative data analysis

To facilitate the computer-aided analysis of qualitative data in the humanities and social sciences and especially in ethnography and sociology, specialist software is continuously evolving that can undertake some of the same ‘stylometric’ analysis done by textual analysis programs but that can also do much more, with more varied documentary sources, using a rather different approach. Whereas textual analysis software looks for ‘strings of characters’ in a collection of documents, programs for qualitative data analysis use a ‘code and retrieve’ system. Codes or tags are added to the documents to denote all sorts of networks, relationships, synonyms and hierarchies as the researcher sees fit. Searches can then be made on the codes and the mark-up as well as on the text of the document itself. Not only can one search for passages where the word ‘male’ precedes ‘authority’ it is also possible to find passages where a discussion of the concept of authority precedes or follows a section on males, men or boys even when the keywords do not appear. Thus, rather than identifying patterns of words, this software allows the user to search for concepts, categories and ideas that have been identified by the researcher. This type of ‘code and retrieve’ program fits with the sort of iterative process that lies at the heart of conventional historical research: reading, querying and hypothesizing.¹⁴

(f) Spatial and mapping software: GIS

The construction of cartograms and other mapping tasks have been revolutionized in the last three decades by the introduction and spread of cartography packages and of Geographical Information Systems (GIS). GIS enables vast amounts of statistical data to be stored, assembled and viewed in relation to its spatial distribution. Several sets of data can be simultaneously mapped according to their geographical coordinates. This enables theories concerning growth poles and geographical concepts such as central places to be put to the test.¹⁵ Major advances in what we might call 'spatial history' have come with the growing sophistication of Geographical Information Science (GISc) since the 1980s. Historical Geographical Information Systems (HGIS) have been developed alongside contemporary applications. This is now one of the fastest growing and most popular areas of the history by numbers, partly because it can incorporate the mapping and analysis of both quantitative and qualitative data.

GIS allows the user to store retrieve, visualize and analyse data that are georeferenced to a location on the Earth's surface. The qualitative and quantitative data relating to a location (*attribute data*) are linked to that location, which might be a fixed point, a line, a polygon or a pixel (arising from the *spatial data*). Initially GIS could only cope readily with quantitative attribute data but as computers have become able to store unstructured texts (books, reports, newspapers, for example), still and moving images, and even sounds, this has opened up the possibility of a much wider array of research amenable to GIS. The ability of GIS to integrate various sets of data with respect to particular geographical coordinates offers great analytical potential in applications such as relating urban overcrowding to the incidence of various diseases or agricultural output to the spread of transport infrastructures, for example.¹⁶ And the extended ability of computers to store qualitative data of all kinds has opened up a whole new field of 'humanities GIS' in recent years.¹⁷ For more guidance on using GIS in a research project, there are some very useful primers available.¹⁸ Google's GmapGIS is available online (gratis) and an excellent software for most purposes. ArcGIS is perhaps currently the most popular with historical geographers and archaeologists (who are increasingly users of GIS).

Most big HGIS projects started a few decades ago with a framework of historical census and vital registration figures so that HGIS was originally associated with historical demography in particular. Many projects have subsequently built upon the spatial demographic evidence overlying it with other sorts of information both quantitative and qualitative as with 'A Vision of Britain through Time', for example (www.visionofbritain.org.uk/). This HGIS database covering the period 1801–2001 includes population and industrial census material, historical maps, electoral results and historical descriptions of places from the pens of travel writers and others. It is linked to the main Great Britain Historical GIS that has reached out beyond history to provide mapping tools for contemporary as well as historical medical researchers and environmental managers.¹⁹ Similarly the US National Historical Geographic Information System (NHGIS) provides, free of charge, aggregate census data and GIS-compatible boundary files for the United States between 1790 and 2014. Other national bodies have followed suit in having similar

History by Numbers

foundational GIS sites and tools: Belgian Historical GIS; China Historical GIS; Portuguese National GIS, for example. In addition many digital mapping sites exist such as the Institute of European History Map Server that contains digital maps of Germany in particular.

HGIS has been employed in a variety of sorts of research projects from mapping agricultural outputs in relation to transportation improvements, voting patterns in relation to a host of potential explanatory variables, the geography of violence, and urban sites of cultural experiment, including film-making.²⁰ An indication of what can be achieved using GIS in relation to the study of medieval wastelands is provided by a case study of County Durham. The research uses Durham Priory charters, rentals, land registers, surveys and accounts, all of which have been entered into an Access database. The evidence from the various charters in the database is additionally converted into a distribution map using GIS. The evidence from the charters contains precisely locatable grants of waste, the overall pattern of which is impossible to see from the piecemeal evidence of individual sources or from enclosure records. By mapping the waste in this way and relating it to physiographic regions, to the later extent of moorland farms and to different types of land ownership (as indicated in Figure 9.4), the authors are able

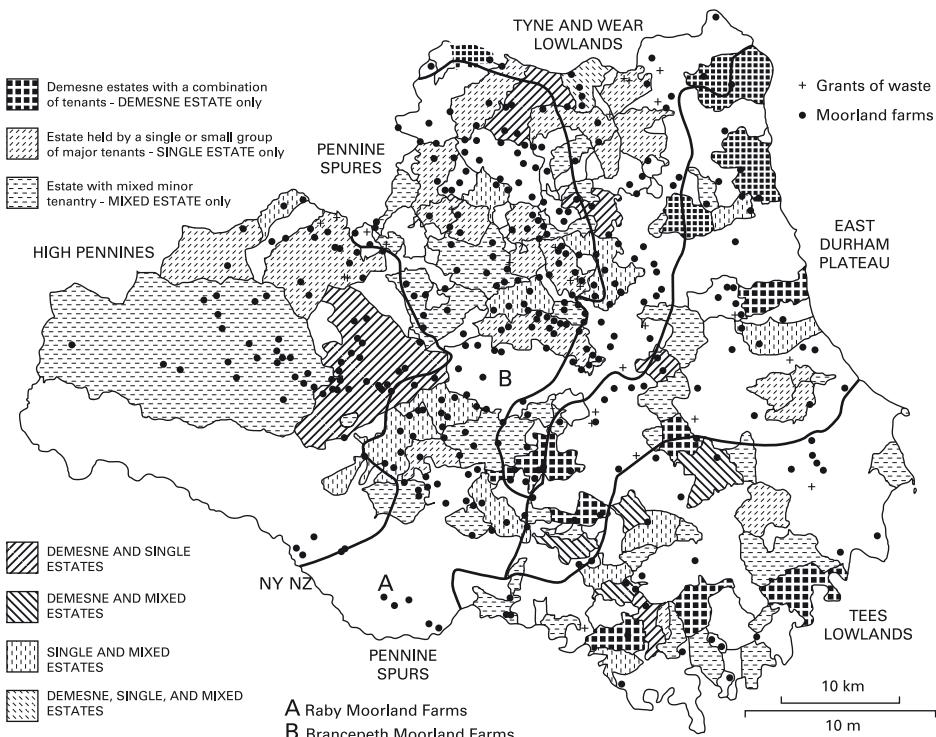


Figure 9.4 Land ownership and grants of waste, County Durham, 1100–1400.

Source: H. M. Dunsford and S. J. Harris, 'Colonization of the wasteland of County Durham, 1100–1400', *Economic History Review*, 56 (1), (2003), pp. 34–56, p. 47.

to demonstrate the proliferation of tracts of waste throughout the period and thus to challenge established notions about the limits of agricultural expansion in the thirteenth century.²¹

In a pioneering study of lay lordship and wealth in landholding in the first half of the fourteenth century Bruce Campbell and Ken Bartley similarly relied upon GIS mapping both to display and to analyse their results at both regional and national levels. GIS enabled them to bring together two important sources: the Inquisitions Post Mortem (1300–1349) and the three major Lay Subsidy records (1327, 1332 and 1334). Mapping these together showed the distribution of wealth and the density of tax payers. Per capita wealth is also mapped indicating the richness of Kent and many parts of the Midlands.²²

The impact of advances in ICT on historical research and writing

Before marrying the statistical techniques covered in earlier chapters with the ability to mount and sustain a piece of historical research, it is important to consider the ways in which advances in ICT have influenced historical research and writing. The increasing sophistication of ICT in the last half-century has had several major influences that are wise to bear in mind when reading computer-aided research (whether highly statistical or not) or when undertaking such research for yourself.

- (a) It has increased the efficiency of archive administration and of data storage and retrieval so that documents and data relevant to a particular research project can now be located and used more easily.
- (b) It has enabled the creation and use of national and transnational databanks and ‘Big Data’ sets that store massive amounts of comparative economic, demographic and social data from across the globe.
- (c) It has simplified and speeded up many old-established approaches and research methods and has allowed them to be applied more extensively.
- (d) It has altered the way in which research is written up and presented.
- (e) It has created the possibility of new sorts of historical research that, in practical terms, would be impossible without computers.
- (f) It has encouraged the growth of particular types of historical research and particular ‘languages’ of historical discussion which have probably been at the expense of others.

Let us consider each of these briefly in turn.

(a) Archive administration, storage and retrieval

Massive increases in the amount and in the accessibility of electronic information in recent decades has transformed the availability of archive lists, bibliographies, press ‘clippings’, microfilmed materials and abstracts from repositories, libraries and other institutions, public and commercial, located all over the world. America led the way in

History by Numbers

encouraging online availability of archive and library resource lists. Britain and other countries are catching up. British developments have been facilitated by standardized rules of archive description pioneered at the University of Liverpool and described in successive Manuals of Archival Description (acronym MAD!) published in the 1980s.²³

Apart from the new ease with which one can gain access to archive lists and bibliographies, documentary sources are often themselves transcribed or digitized into machine-readable form and their contents can be called up on screen. Much material has been scanned, not always as carefully as it might be. There is often no substitute for dealing directly with an original source, although the temptations to avoid this and 'make do' with electronically available material is great. Often sources, such as large format parchment deeds for example, do not lend themselves to efficient or effective digitization whilst large-scale digitization projects, of newspapers for example, often encounter the technical limits of rapidly scanning text where documents are either not printed clearly in the first place or have suffered from deterioration over time. Indeed transcription and digitization are often driven by conservation considerations which tend to bias what is available. In other cases commercial publishers have engaged historians to advise on appropriate archive collections and selections (generally printed ones) that are then digitized. Here a commercial selection is made; preference is often given to printed, rather than handwritten, sources because these are easier to scan. There is also the inevitable problem that all collections of archive material available on the Internet are selective and partial. Commercial considerations come into play but much influence is also exercised by those historians paid to advise upon which documents to include. Often this advice is based upon their personal research agenda and their future ambitions. Another pitfall of the spread of such easily available primary source material is the danger that researchers will take the easy course of prioritizing research that uses this data and the archives which are more difficult or slowest to be transcribed or scanned will be relatively neglected. In defence of these developments, the ease of availability of important research collections has 'democratized' many subjects of study. Many important collections available previously only in cities housing major libraries can now be researched easily by academics, and others, living more remotely and/or in various parts of the globe, and by disabled as well as able-bodied scholars. This broader catchment of researchers stimulates increased debate on a global scale, facilitated in turn by electronic networking and electronic discussion groups.

(b) Databanks and Big Data

Taking the availability of historical evidence a step further, material from original sources has frequently been processed by researchers to form many, often very large, databases, spreadsheets or marked-up text archives which are then deposited in databanks and the files are made available to other historians. For example the Economic and Social Research Council (ESRC) has a large computer-data storage facility at the University of Essex where researchers who have had ESRC grants (and others) are encouraged to store their data with Internet access so that others can use it.²⁴ The Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan

similarly holds machine-readable data archives for 350 member institutions. These are strongest on election and census data which cover 130 countries (to varying degrees).²⁵ Similar large social science and history collections can be found in the Netherlands, Norway, Denmark and Sweden whilst large and growing text archives are located, amongst other sites, at the Universities of Oxford, Rutgers, Groningen, Pennsylvania, Pittsburgh and Washington DC.²⁶

An increasing amount of research, especially in economic, demographic and social science history is now done by historians exploiting the possibilities of such **Big Data** deposits. Big Data generally refers to resources of quantitative (and other) information, stored electronically and often amounting to several terabytes. Big Data often draws comparative evidence from many parts of the world. The volume and diversity of data make it difficult to store, retrieve, analyse and utilize. Indeed whole degree courses have now been developed to train experts in this field.²⁷ As governments as well as academics and health professionals will soon be routinely using petabytes of data stored in multiple formats across different platforms, experts with the skills and knowledge to design and deploy complex information systems are badly needed. For historians some useful guides to current practice and potential have been assembled, most notably Patrick Manning's *Big Data in History: A World Historical Archive*, which introduces a project at the University of Pittsburgh that aims to create a world-historical archive tracing the last four centuries of historical dynamics and change.²⁸ The Big Data archive at Pittsburgh links research on social, economic and political affairs, health, and climate, for societies across the globe. The advanced technology and attention to detail that go into building such an archive and the pitfalls and benefits (mostly the latter!) of gathering and disseminating data for historical research across five continents are mapped out in Manning's book.

One area of Big Data use in recent decades has been in comparative research on economic growth and development using GDP or GDP per capita. Sometimes wider comparative measures are used, such as the Human Development Index (HDI). We have seen some examples of this earlier in the volume. Such projects often involve teams of scholars working in different national contexts. Much of the resulting historical data is housed at the Rijksuniversiteit, Groningen as part of the Maddison Project, named after Angus Maddison whose research pioneered the use of historical national accounts for comparative purposes.²⁹ Another growth area has been international collaborative census research such as that carried out at the University of Minnesota Population Centre (using the Integrated Public Use Microdata Series IPUMS-USA and IPUMS-International).³⁰ The Mosaic Data Archive project currently located at the Max Planck Institute in Rostock 'identifies, gathers, harmonizes, and distributes surviving historical census microdata for Europe and Beyond' including the latest British historical data generated at the Cambridge Group for the History of Population and Social Structure.³¹ The Mosaic project builds in turn upon various integrated census projects such as IPUMS and the North Atlantic Population Project (NAPP) which includes the complete censuses of Canada (1881), Denmark (1787, 1801), Great Britain (1881, 1911), Norway (1801, 1865, 1900, 1910), Sweden (1880, 1890, 1900), the United States (1880) and Iceland (1703, 1729, 1801, 1901, 1910).³²

History by Numbers

There are of course problems as well as advantages with these Big Data collections. First, the researcher will be relying on information extracted by someone else who had different interests or priorities. Any errors in the original transcriptions will be repeated and perhaps magnified in the course of further manipulation; information omitted from the original transcriptions or scanning may get forgotten and there may develop a tendency for historians to analyse the same body of data again and again (with diminishing returns), rather than seeking new or revised evidence. It would certainly be a great mistake to carry out research entirely without contact with the original records and their wider context in other documents and sources, not so easily amenable to computer storage or analysis. In addition, as we have noted, many Big Data projects involve the use of comparative data from different parts of the world and originally collected by different institutions or nation states using different methods, categories and criteria, which the historian must recognize before they can be reliably compared. Some Big Data projects devote much time to making various transnational datasets comparable or ‘integrated’ but the problem remains a major source of criticisms of comparative studies of economic growth using GDP and GDP per capita figures collected across the world by different political and economic entities. Where studies involve time series, relating particular growth indices, demographic characteristics or other social or political evidence, extending over centuries, the problems of comparability, of comparing like with like, multiply.

(c) Speeding up old-established approaches and research methods

This is a particular advantage and feature where research involves cumbersome and time-consuming statistical applications but it is also very apparent in any research involving the storage, retrieval and manipulation of large datasets. Projects are now undertaken which would have been simply too time-consuming and hence too costly to contemplate without computer use. This does have a downside however. It is now so easy to run correlation and regression software that it has had the detrimental effect of blinding some historians to the shortcomings and inappropriate nature of the data for such purposes. The ease with which statistical analysis can be accomplished can also lead to very poor research practice. For example, the historian should always have a very good reason for posing the possibility of causal connections between variables before undertaking relevant statistical analysis. A hypothesis should be framed at the outset which is capable of being falsified. Poor practice creeps in where many regressions are run using different combinations of variables until a ‘significant relationship’ (statistically speaking) turns up. The historian then concentrates upon analysing the ‘cause’ of such a relationship, ignoring the fact that the correlation may be spurious or accidental.³³

(d) Writing and presenting research

Word processing, and the interface between word processing and statistical and other softwares, has altered the way in which history is written up and presented. There is a greater tendency now to use graphical and other visual representations because they are so much easier to prepare and to incorporate than in the past and because they often look so much more impressive than formerly. There is little doubt that tables, figures and

graphs provide the most concise way of conveying extensive or complex data and that they have the potential to clarify arguments. However, great care must be taken in using them appropriately, as the earlier chapters of this book have pointed out. Sometimes, the easy formation of figures, graphs, cartograms and so on are made to substitute for careful thought about what is necessary or sufficient, most appropriate and least misleading.

The impact of word processing upon the nature of creative writing across a range of academic and artistic activities is thus far little understood although it is likely to have been significant and will no doubt have its researchers in the future.³⁴ The current emphasis upon language and vocabulary, narrative and rhetorical structures of academic works (as well as historical sources) which has been influenced by post-structuralism and by other linguistic theories, such as that attributed to Mikhail Bakhtin, suggests that attention to the way in which things are written will only continue to grow.³⁵

Do people write differently when they use a word processor? The ease with which paragraphs and pages, quotes, inserts can be moved around in a text must inevitably have had an impact on the way in which history (along with other subjects) is written. Word processing can be an uninhibiting release from the tension that used to surround 'putting pen to paper'. With handwriting or typing there was always the thought that a first or early draft must be near-final because of the time-consuming and costly problems of multiple redrafting. Computer use may encourage drafting and redrafting until better pieces of work are produced. Or it may encourage laziness in transposing sentences and paragraphs from one draft or publication to another. The production of camera-ready copy and computer typesetting may also encourage both authors and publishers to publish prematurely and without the rigorous discipline of checking which the older system of galley-proofs demanded.

The fact that most writers now use word processing will at the very least mean that less evidence of the writing process is likely to survive. Marginal notes, improvements, alterations from one draft to the next are lost because only the discs containing the final versions will probably survive (if those) for the historians and literary theorists of the future.

(e) New sorts of historical research

The biggest recent growth areas of computer use in new sorts of historical research have involved relational database software, spatial mapping tools (GIS) and software that enables the storage and analysis of unstructured texts (opening the way to greater use of computing with qualitative data). As we have seen, relational database software enables multiple files of information to be built up and then linked together or cross-matched to create new features of historical evidence. This can be seen most clearly with the development of family reconstitution. Files of information are created from baptism, marriage and burial data from parish registers. Software is then developed to instruct the computer to match up these entries on the basis of carefully thought-out linkage criteria (algorithms), so that families can be 'reconstituted'. Before computer use, although the technique had been developed it was only possible to apply it to very small communities because of the time taken to enter data on cards and to link records by hand, rather than

History by Numbers

'automatically'. 'Automatically' is here placed in inverted commas because although linkage algorithms may be written and employed there will remain many cases where the linkage criteria either fail to secure a potentially correct link or where the criteria result in mistaken links and only hand sorting at the margin and spot checks can go some way to counter these difficulties.

Reconstitution can be taken some steps further in a more fully developed community reconstruction study that might involve data linkage across files of demographic evidence, tax records, wills, inventories, local government records, business and estate papers and so on. The same individuals often figure, momentarily at least, in different documentary sources. Through linking data by surname, forename and other identifiers (that is, by nominal record linkage) new sorts of detailed research on the lives of otherwise relatively unremarkable people can be carried out. From the accumulation of many small shreds of evidence, a picture of the life-course and lifestyles of ordinary people can be glimpsed. A project illustrating this sort of research is a study of small business families in north-west England between 1760 and 1820 conducted by Hannah Barker and Mina Ishizu at the University of Manchester, the results being available on a publicly searchable database.³⁶ This new sort of 'history from below' would not be possible without computers because it would be impossible to both store and link shreds of evidence across such a large number of separate sets of information, using index cards and by hand. One could argue that computer-aided research is rejuvenating local and micro-level history, freeing it from its antiquarian roots and providing useful alternative insights to those provided by macro-level studies.³⁷

Similar projects to reconstruct information about local and non-local groups (such as professional communities), have been undertaken, for example the study of early modern medical practitioners at the University of Exeter and the Legacies of British Slave Ownership Project at University College London which builds upon a database derived from the slave ownership compensation scheme that operated after the abolition of slavery in the British empire in 1833.³⁸ The Westminster Historical Database project at Royal Holloway and Bedford College is another good example. It started with digitization of the Poll Books but is now linked to the much larger project, 'London Lives, 1690–1800: Crime, Poverty and Social Policy in the Metropolis', a publicly available searchable database that enables searching over 3.35 million names derived from 240,000 manuscripts across eight archive sources and fifteen datasets.³⁹

GISC has promoted the application of several datasets in relation to the same spatial coordinates enabling the analysis of such things as housing, health, incomes, voting patterns and so on to be mapped onto one another. Thus on the website 'A Vision of Britain through Time', travellers' descriptions of parts of Britain can be mapped onto conventional statistical sources for the same geographical areas to produce a hybrid understanding of conditions and of change. Historical urban topographies can also be precisely reconstructed in terms of social as well as geographical coordinates, architectural patterns and literary or cultural manifestations whilst political violence can also be subjected to social and geographical environments on the ground. Good examples of these areas of research are (respectively) the major project reconstructing the social

and architectural landscape of Kyoto in the Edo period; new research on amateur and professional films made in and about Liverpool in the twentieth century and a recent GIS-based study of political violence in Northern Ireland during the ‘Troubles’.⁴⁰

Many new sorts of historical research that can be undertaken using machine-readable data rely upon the possibilities that emerge once a vast amount of information (quantitative and qualitative) has been digitized. For example, Jane Humphries’ major study of child labourers during the Industrial Revolution was made possible by her ability to link conventional quantitative sources with a mass of individual detail from more than 1,000 working-class autobiographies. Thanks to this we now have a much more representative and reliable picture of the age of starting work, conditions of work and wages, and the contribution of children to household incomes.⁴¹ Perhaps the biggest and most original new area of research in economic and social history that has been made possible by computers is the study of living standards and nutrition using anthropometric data.⁴² Only through the inputting (hence machine readability) of mass evidence on heights and weights from manifold sources (from prison records, convict surveys, military listings, school or medical records, for example) can data on human measurements be sufficiently dense and representative to reveal patterns about wider experience in the entire population. Such data has necessarily to be combined with knowledge (from the health sciences) of the impact upon adult heights of nutritional levels, including nutritional insults, at different ages and stages of development. Often a mass of skeletal evidence is used.⁴³ We encountered some debates in anthropometric analysis earlier in this volume (pp. 179, 189). This sort of research includes debates about the impact of various diseases, such as smallpox, occupations, such as coal mining, and periods of famine, in stunting adult heights.

Another area of research just about impossible without computers is the intensive textual analysis of large bodies of documentary sources. Machine-readable prose sources such as speeches, broadsheets, pamphlets, biographies, diaries, novels, oral history transcripts, can be analysed using textual analysis mark-up and software programs to allow word frequency counts, word associations and common phraseology to be identified, as well as to speed up analysis of content. These techniques have been used in verifying the authorship or authenticity of various documents as well as in analysing changes and constants in language and vocabulary use. As software for the latter purpose becomes increasingly sophisticated, as more corpora and prose-source material is digitized and as history continues to focus on the importance of language and discourse, the use of this type of computer analysis is likely to grow.

A recent example can be found in Phil Withington’s study of the changing meaning and use of the word ‘peace’ in England between 1500 and 1700. Through a series of simple word counts of ‘peace’ and ‘war’ from a number of bodies of text, most notably Shakespeare’s corpus, Early English Books Online (EEBO) and the English Short Title Catalogue (ESTC), Withington is able, alongside non-quantitative evidence, to suggest the increasing use of ‘peace’ to legitimize and sanction violence after 1640, stemming from its role in describing society and the self as well as spiritual and civil life.⁴⁴ Figure 9.5 illustrates a further example of word analysis, using corpus linguistics. It shows the

History by Numbers

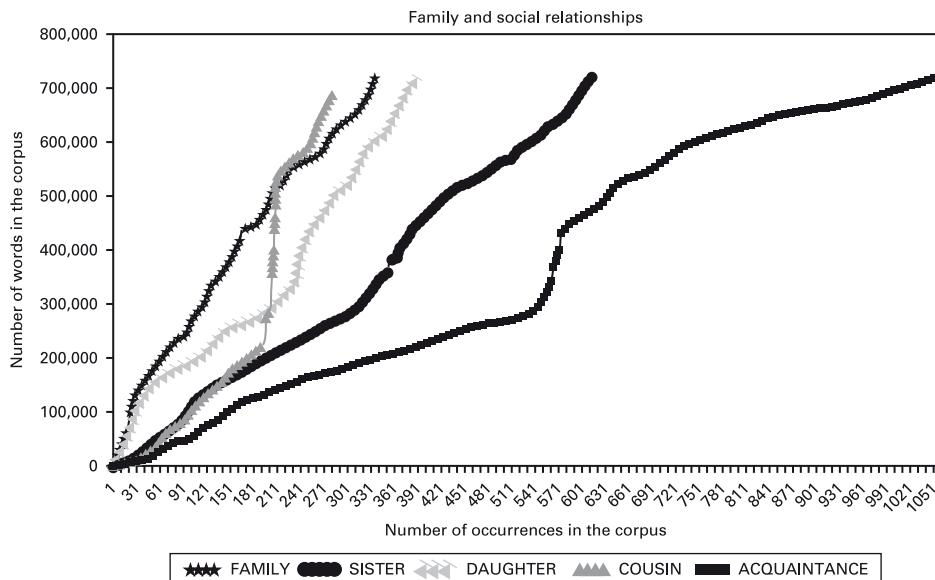


Figure 9.5 Family and social relationships in Austen.

Source: Bettina Fischer Starcke, *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries* (London 2010), p. 178.

occurrence of various familial words in Jane Austen's work enabling the author to say much about perceptions of family in early nineteenth-century polite society.

Perhaps the biggest future growth area for computers in history will be the use of software, not just for the stylometric analysis of relatively homogeneous bodies of text but for the storage, retrieval and analysis of heterogeneous documentary evidence derived from primary source transcripts or scans, information arranged in matrices together with the content of tags and codes that incorporate the historian's own thoughts on the evidence. The use of computers for such qualitative data analysis, especially in history, remains in its infancy. A major disincentive to the spread of such techniques is that it can only be justified for certain projects where the source is not too extensive, and where the text promises to yield coherent 'webs' of coding. The amount of time required at the outset for text mark-up, annotating and coding, classifying and categorizing is a further disincentive but once these are done the potential of computer-aided analysis is great.⁴⁵

(f) Encouragement of particular types of research

The use of computers in all fields of quantification and statistical analysis is inevitable because they facilitate and speed calculations and produce accurate and neat diagrams and figures. However, they are also attractive because their use endorses the idea of scientific endeavour and objectivity. The freedom from subjectivity that is assumed to come from applying fixed rules and procedures to calculations lies at the heart of the idea of statistics as the language of science. Computers allow the application of the

fixed rules of statistical analysis and inference with the least danger of subjectivity and personal intervention. As we discussed in Chapter 2 this is not always a blessing: it is simultaneously an advantage and a disadvantage in historical work. Disadvantages multiply where statistical software makes numerical analysis so easy that it can be done by those without proper and full understanding of the concepts and pitfalls involved.

It is certainly the case that early computer use in history encouraged quantification and quantitative methods of all kinds, including econometric history. For a time, in some branches of history, computer-aided cliometrics was seen as the high-status cutting edge of history whilst other approaches became less valued, especially in economic history. One of the problems was that early software limited computer use to highly structured primary sources, such as censuses and tax returns, export and import figures, output statistics and population totals, and to a restricted range of largely quantitative research questions. Although software and hardware developments now mean that computers are no longer associated solely with a narrow range of sources, approaches or techniques, there is still the probability that computer use is leading history in certain directions and encouraging some approaches and methods, and the study of certain time periods and topics, at the expense of others.

The digital revolution of recent years has resulted in the easy availability of trillions of words of printed text.⁴⁶ It has been argued that the somewhat arbitrary and varied technical nature of the build-up and design of digital databanks, often influenced by commercial considerations, has at worst resulted in a sort of 'research roulette dressed up as traditional scholarship'.⁴⁷ The digital revolution has also created selection bias in the sources that are available to use. The relative paucity of digitized non-Western sources has led one historian to suggest that the process of digitization itself has endorsed a Western hegemony in ways that would not be acceptable as the product of a more self-conscious policy.⁴⁸ Historians working on British history in the early modern period and the nineteenth century are perhaps the beneficiaries of the most thoroughly digitized place and period in history.⁴⁹

As new software suitable for historical applications is developed, research runs the danger of being biased in favour of the new techniques and approaches involved. The earliest products for textual analysis encouraged a focus upon relatively homogeneous bodies of text and overemphasized collocations and keywords at the expense of rhetorical and narrative structures and other aspects of text which were more difficult to identify and measure using computer-aided resources. The new qualitative data analysis tools of the last few decades, whilst promising much for histories of the future, also run a risk of biasing research in favour of certain sorts of sources and issues that lend themselves best to the technique. More problematically, this approach foregrounds the iterative process whilst also reducing the separation of theory from data upon which this rests. If the historian's division of the data into overlapping 'data-bits', her notes and codes all become part of the dataset and merge with the primary evidence itself, iteration and especially the formation and testing of hypotheses may need to be replaced by something more akin to 'thick description'.⁵⁰

Getting started with quantitative and qualitative historical research employing ICT

There are two basic ways of starting out on a piece of research or a dissertation, as outlined below.

Problem-oriented approach

The problem-oriented approach arises where the researcher starts with a subject area that interests him or her and identifies research questions, hypotheses or issues from the secondary literature. For example, one might start with the construction of gender in the nineteenth and early twentieth centuries as the broad area of research, narrowing this down to the role of work and work relations in constructing gender ideas and ideals. A source (or sources) is then found (perhaps employment records, autobiographies or diaries, works' newspapers or the records of the Mass Observation movement (see p. 54) that helps to address these questions. This may also help to pin down a finite period that your specific case study or piece of research will address. You may initially have a broad topic in mind but for primary research purposes, it is necessary to hone this broad interest down to something that is do-able in the time period that you have both for the initial research itself and for the writing up.

Let us assume that the researcher has decided that (quite an unusual and indirect source) folk songs and ballads of the period (which often have an occupational or industrial context) may reveal something of the relationship between gender and work. Perhaps the researcher then chooses to concentrate upon agrarian, seafaring and mining songs. There is much secondary literature available on the history of labour in these sectors: this will give a rich context to the research as well as raising the questions that might be asked. A comparative study is often valuable in underpinning both research questions and results. The bearing which these sources have on other issues may be largely ignored for the purposes of the dissertation at hand. It is important not to get sidetracked.

As one becomes familiar with the source, it may be necessary to adjust the questions asked in light of what the sources seem able to yield, for example the clustering of songs around dairying and whaling, and around mining disasters may suggest that the research will focus on these. We could then go on to employ a database or textual analysis package to reveal quantitative data that demonstrates the potential of the source for revealing certain gender issues: perhaps the gender-based division of labour, perhaps the differential impact of gender in industrial unemployment, disputes or accidents will come to the fore.

In another example we might decide to research an aspect of trade between Britain and India in the eighteenth century. Perhaps one of the following hypotheses (arising from the secondary literature) might be pursued: that the role of patterned cotton textile imports from India was central in driving wider trade or that the importance of certain mercantile and family networks in developing aspects of commercial exchange or inspiring new tastes for Indian goods was vital. Perhaps we decide to use a subset of the East India Company records of voyages, imports and exports or perhaps the private papers of certain individuals involved in settling, carrying out military operations and

establishing trade links. Business records from the East India Company might involve us in the creation of a database and spreadsheets relating to specific voyages involving the importation of Indian cotton cloths. Private letters and similar documents might be analysed in a traditional and non-quantitative way or one might choose to quantify as well as to analyse networks between merchants and families using diagrams, tables or graphs derived from the rearrangement of selected information into a spreadsheet or (if digitized sources are available) by employing textual analysis software.

One difficulty with the problem-oriented approach, as these two examples suggest, is deciding which sources are appropriate and most important to the initial interests and issues posed, and where to draw the line in collecting evidence. In both our examples we might decide (perhaps when it is too late if a finite piece of research producing results in a given time period is the aim) that such issues cannot be properly addressed without a huge expansion in the array of data consulted and analysed. A key need in all research projects, large as well as small, is to have a clear view of the necessarily partial nature of any answers revealed by the work. Indeed it is often wise and good practice to allude to further data and analysis that might be used to follow up what has been achieved.

Source-oriented approach

In the alternative source-oriented approach a source or group of sources is uncovered or selected, for example, trade directories, an informative collection of correspondence or diaries, a good set of business records, or the census enumerators' books for an urban area. The content of the source is then allowed largely to determine the nature of the enquiry. For example, trade directories generally give information about the various sectors and types of businesses in a town or region and were published periodically in the eighteenth and nineteenth centuries. Although they often omitted smaller concerns and were sometimes insufficiently revised from one edition to another they can be used to assist in addressing many important questions about the nature and ownership of business in a city, town or region, and about change in such characteristics over time. They can be used to identify changes in female entrepreneurship for example. Diaries and letters can be used to address the nature of familial, social or business relationships, the nature of affection or dispute within families and many other social and cultural phenomena. The choice of precise research topic based upon such a collection of papers will largely be dictated by the extent to which such matters are addressed and some sample reading of the letters or diaries might be needed to check on this. Business records for the Yorkshire worsted industry in the nineteenth century might similarly dictate a study of the way in which a firm kept their records, their use of credit, their fluctuating fortunes, the geographic distribution of their purchases and sales, and their employment and payment histories, for example. Census enumerators' books for a residential area would enable many questions to be asked about household size and structure, migration, occupations, and the role of lodgers and live-in servants.

The major problem with this source-oriented approach, across all of these examples, and others, is that (on its own) it may lead to a piece of history that has no clear focus and

History by Numbers

which spreads itself over too many disparate issues. It will be difficult to discern what is of value without reference to some external set of questions which would normally be rooted in the secondary literature. It would certainly be necessary to refer to studies that have used similar sources and to issues that they have raised. It will also be necessary as the sources become more familiar and the research project more clearly defined, to consult the secondary literature on the particular research questions addressed. All projects, whether primarily problem-oriented or source-oriented, demand a firm contextualization in relevant research and writing that is already available. New research should always be focused upon filling gaps in our knowledge of a subject, or in either challenging or endorsing existing conclusions.

Research in practice

Usually a balance is struck between the two approaches. Emmanuel Le Roy Ladurie's study of *The Peasants of Languedoc* started out as a study of one source, the land tax, and was broadened to an investigation of the whole social structure of the area and demographic change:

Mine was the classic misadventure; I had wanted to master a source in order to confirm my youthful convictions, but it was finally the source that mastered me by imposing its own rhythms, its own chronology, its own particular truth.⁵¹

Often the historian will start with one set of questions or a hypothesis, with a problem-oriented approach only to find that the sources cannot answer them and that they direct the work on a rather different path. Alternatively one can start with a source, find that it will be particularly good for addressing one or two specific questions and realize the need immediately to consult the secondary literature to see what is already known about the questions or what similar work has been done on such sources. Once the secondary literature is mined for contextualization of the evidence, questions and hypotheses assert themselves. It would be foolish to think that one could initiate a historical project which was entirely source-oriented and without preconceived questions or ideas. It is also the case that a problem-oriented approach must be ready to modify the original objective in the light of questions that may arise directly from the sources. In practice research is an iterative process. It might be initiated by one or other of the approaches detailed above but the result will arise from journeying back and forth between the two, and the related secondary literature, until the project settles down with defined questions and a finite scope.

Research projects and ICT

Early in a research project or dissertation the question of whether to use qualitative or quantitative research methods, or a mixture of the two, will arise. Along with this,

the extent of ICT employed will also arise. Initially this might be determined by the primary source material that may be available, not in an archive or in the field but online in either digital or database format. Further computer use will depend upon the type of source material and the amount of source material. Computers can be a boon if there is a relatively large amount of easily classifiable data that needs neatly summarizing and/or analysing statistically. Figure 9.2 shows a page from a land tax return for Yorkshire. This is a good example of a semi-structured source which is relatively easily adapted for computer-aided descriptive statistics: the information is already in columns and if several land tax returns are to be compared the distribution of landed wealth (as reflected in tax paid) can be assessed and studied for different areas or over time. As is always the case, a close knowledge of the nature of the source is required alongside the ability to manipulate it. In the case of the land tax, particularly if one is interested in changes in the distribution of wealth over time, one must be aware of the fact that the schedules often went unrevised from one year to the next and a study of landholding and land ownership would really need to be supported by other sources such as estate records, surveys or tithe returns. One could nevertheless make some tentative additions to knowledge by constructing frequency distributions and pie charts reflecting the distribution of tax paid as we did in the case of some of the Sowerby figures (as an example) in Chapter 3. One could use a spreadsheet or database here or move the data from one to the other. If one wanted to look for the same names appearing in different tax returns, one could either create a table for each return and use relational database linkage *or* one could input all the returns required into the same database (making sure that unique identifiers and year fields are present for each case). Either method would enable one to test whether the tax paid by particular individuals had grown or declined over time and one could probably imply from this some constants and some changes in the distribution of holdings.⁵²

Another source that one could use to create a spreadsheet and undertake some interesting statistical work would be parish registers. One could use aggregate figures monthly or yearly from the three classes of so-called 'vital events': baptisms, marriages and burials, forming a matrix from these in a spreadsheet program. These could then be used to test for similar movements over time through the construction of simple graphs, histograms or scatter graphs. More involved inferential statistics may also be appropriately applied, particular time series and growth rate estimates and projections, and correlation and regression analysis. The data displayed in Table 5.10, for example, contained baptism and burial figures for St Martins, London. Plague deaths which were noted in the source have also been recorded. Descriptive summary statistics could be applied to such data with benefit, the relationship between baptism and burial trends could be investigated, growth rates of each calculated and the periodicity of any cycles of fertility or mortality (perhaps associated with plague) might be identified, especially if we had the much longer run of the data from which Table 5.10 has been extracted.

Alternatively, detailed data from each entry could be tabulated in a database containing three tables, one for each set of vital events. For a very small parish or chapelry, it might be possible, in a dissertation, to input all events over a sufficient period of time to undertake some elementary family reconstitution. One would not need a semi-automated

History by Numbers

program or advanced algorithms for this (such as used by the Cambridge Group). The ability of a database program to sort and retrieve would enable some computer-assisted, manual reconstitution of families to be done along with associated calculations about average age of marriage, death, family size.⁵³ The occupational information contained in many parish registers together with details of illegitimacy enables further interesting questions to be addressed.

When contemplating analysis of a text or texts: speeches, folksongs, newspapers, autobiographies for example, the question of whether to use software for textual or further qualitative analysis arises. An important consideration here often is how much text would need inputting (and with what level of mark-up) because this is what will take the most time. More important, of course, is the question of whether the documents are appropriate for the application of such analysis. Speeches, political tracts and novels have proved most adaptable to 'stylometric' analysis. Autobiographies and diaries are likely to be other appropriate sources though only a limited amount of this sort of work has been done on them to date. Huge collections of machine-readable prose: novels, journals, letters, diaries and other papers are now available online, following digitization projects of various kinds and sizes worldwide. Most are open to others to use for their own analysis. And much can be achieved with the sort of simple nominal and other word searches normally built in to the digitization that has been provided as with the Old Bailey Court Reports (www.oldbaileyonline.org/) and the various databases in London Lives (www.londonlives.org/). Indeed the Old Bailey records have been used as the basis for many different sorts of research beyond crime itself, including gender, master-servant relationships, time use, and the expansion in ownership of consumption goods (using the records of stolen items), as we have seen (p. 218).

For reasons of time and convenience student projects, whether prose-based or using statistical data (or both), often employ existing machine-readable files instead of creating new ones. Much data is available online from the ESRC data Archive at Essex and from similar databanks in the United States and Scandinavia in particular.⁵⁴ A ready-made spreadsheet of information, a database or a machine-readable set of texts may seem attractive and may circumvent the need for laborious inputting of data but they do also create problems for the new researcher. Despite the advantage of saving time and effort in inputting, it must be recognized that difficult decisions about the structure of the tables and matrices, categories and codes have generally already been made, geared to other purposes or standards. The choices and decisions of previous researchers may distort the application of the data for new research purposes. This is an old problem as the same difficulty applies to all information which has been collected and tabulated for purposes other than those of the researcher. The additional problem in this case is that the data may have been processed several times by the time it is preserved in a machine-readable data archive. It is always a good idea to ask the following questions and to discuss their implications in the introduction to the dissertation:

- (a) What do we know about the nature of the original documents and the ways in which elements of evidence, layout, style, form as well as content, might have been obscured by the 'translation' to machine-readable form?

- (b) What other sources might one have consulted if they had been as easily available as the machine-readable data?

Things to look out for when embarking upon a computer-aided piece of historical research (whether quantitative or not)

- (i) Always select a dissertation or project topic of manageable size particularly in relation to the amount of primary source material that you intend to analyse. Draw up a schedule to fit the time available allowing time for emergencies and delays. Carefully assess the time it will take to input data into the computer (time the entry of each row of a database, for example, or of each page of a text and multiply this by the number of cases or pages you will be including). Never underestimate how long the dissertation will take to write up. Whatever your estimate for writing-up time, you would be wise to double it.
- (ii) Your choice of software will usually be a compromise between what would be ideal for your purpose and what is available on the server at your institution or on your PC. You will usually have some simple spreadsheet applications, database programs, text analysis software, sufficient for most common purposes. More advanced software and tools might be explored with advice from your home institution, although introductions to most software can be found on the Internet (see online Appendix). Sophisticated software for qualitative data analysis, spatial and textual, is also available and there are Internet guides to their use. The first problem with all of these, however, is that none will have been designed specifically for one's own project and most will not even have been designed with historical research or sources in mind. Some of the most difficult decisions one has to make concern which package to use. Whether spreadsheet or database, for example, and whether you will need an additional graphics and statistical interface. Software availability is changing all the time and institutions vary in what they have to offer on mainframe servers. The advice of supervisors or colleagues is often going to be critical.
- (iii) If using a spreadsheet, think carefully about the structure of the matrix and the content of the rows and columns. If using a database, spend time perfecting the database design (especially the number, type and size of fields), so that you will be able to include all the information which you need for your analysis, and make sure that the characteristics of field size and form are the same across any tables that you intend to use in a relational structure.
- (iv) If a field of information needs to be coded for analysis think hard about doing this and, if possible, leave it until after the inputting stage, adding an additional database field for the codes. You can then always go back to the 'original' entry. When coding nominal files such as occupational ascriptions be sure to use appropriate guides to nomenclature. Standard occupational classifications are often appropriate but can be misleading if applied to very different time periods or cultures than those for which

History by Numbers

the classification was established.⁵⁵ In the same way it is always best to input all the variations of spellings that occur in the original document. Fuzzy searching minimizes the problem of missing linkages and inaccurate searches that arise simply because of variations in spelling. It is usually only possible to know when a variation reflects a real difference long after the inputting has begun, by which time it may be too late to resurrect the original data with any ease. Always retain as much as possible of the richness of the original fields of information when forming a database.

- (v) If inputting a text for textual or qualitative analysis make sure that the mark-up conventions and procedures are fully understood from the outset and consistently applied.
- (vi) Finally, all dissertations and projects should include a historiographical context, a set of justifications for undertaking the research, discussion of the advantages and pitfalls of the chosen source material and some discussion of the strengths and weaknesses of the theoretical and methodological approach taken. Always remember that you are in charge and not the data or the computer.

Conclusion

In almost all branches of the historical profession computers have revolutionized the way in which history is researched and written. They have facilitated the location and the analysis of sources and have made many sorts of research projects viable or possible for the first time. There are, however, costs and pitfalls involved. The digital revolution has created a mass of data of varying quality and consistency and the easy availability of machine-readable data often means that the richness of original sources is neglected. The ease of statistical manipulation may result in bad practice and in too little attention being paid to the reliability of the data or to historical context. By initially giving a tremendous impetus to quantitative work of all kinds computers have been responsible for aggravating the division that already existed in the profession between supporters and critics of quantification. Interestingly, more recent software developments are now healing the rift by opening up the possibility of extensive computer use in text- and object-based analysis, in the non-quantitative as well as the quantitative study of communities, and in research using very varied qualitative data. What must however be remembered is that all computing depends on numbers: variable optical character recognition (OCR) by the computer of digits and characters, their manipulation, rearrangement and counting; arranging, rearranging and categorizing data; making connections between categories; forming matrices, adding and averaging; recognizing and displaying frequencies; looking for and measuring associations; calculating probabilities. Computers rely upon all these basic tasks associated with statistical manipulation. Thus, computer-aided research, whether dealing with quantitative or qualitative evidence (or a mix of the two), is always an aspect of 'history by numbers'. As such it is subject to the same debates about the competing values of precision and ease of

communication, on the one hand, discourse, rhetorical argument and narrative on the other. But, providing one remains aware of the pitfalls of all approaches, there is no reason why we must take sides in these debates nor why historians cannot dwell upon concepts, subtleties, ambiguities and the detailed description of single events or individuals at the same time as they are cautiously guided by the ‘science’ of numbers.

Further reading

- Alves, Daniel (ed.), ‘Digital tools and methods for historical research’, Special Issue of *International Journal of Humanities and Arts Computing*, 8 (1), (2014).
- Anderson, Ian, ‘History and computing’ at: www.history.ac.uk/makinghistory/resources/articles/history_and_computing.html (accessed 9 March 2016).
- Bodenhamer, D. J., J. Corrigan and T. M. Harris, *Spatial Humanities, GIS and the Future of Humanities Scholarship* (Bloomington 2010).
- Cameron, Sonja and Sarah Richardson, *Using Computers in History* (Basingstoke 2005).
- Coffey, A. and P. Atkinson, *Making Sense of Qualitative Data: Complementary Research Strategies* (Thousand Oaks, CA 1996).
- Cohen, D. and R. Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web* (Philadelphia 2005).
- Dey, Ian, *Qualitative Data Analysis: A User Friendly Guide for Social Scientists* (London 1993).
- Freeman, Mark, *Quantitative Skills for Historians* (Warwick 2010).
- Graham, Shawn, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: The Historian’s Macroscope* (London 2015).
- Greengrass, M. and L. Hughes (eds), *The Virtual Representation of the Past* (Surrey 2008).
- Greenstein, D. I., *A Historian’s Guide to Computing* (Oxford 1994).
- Gregory, I. N., *A Place In History: A Guide to Using GIS in Historical Research* (Oxford 2003).
- Gregory, I. N. and A. Geddes, *Toward Spatial Humanities: Historical GIS and Spatial History* (Bloomington 2014).
- Harvey, C. and J. Press, *Databases in Historical Research* (Basingstoke 1996).
- Harvey, F., *A Primer of GIS: Fundamental Geographic and Cartographic Concepts* (New York 2008).
- Heywood, I., S. Cornelius and S. Carver, *An Introduction to Geographical Information Systems* (4th edition, Harlow 2012).
- Hitchcock, Tim., ‘Confronting the digital or how academic history writing lost the plot’, *Cultural and Social History*, 10 (1), (2013), pp. 9–23.
- Hudson, Pat, ‘A new history from below: computers and the maturing of regional and local history’, *The Local Historian*, 25 (4), (1995), pp. 209–222.
- Lewis, M. J. and R. Lloyd-Jones, *Using Computers in History: A Practical Guide to Data Presentation, Analysis and the Internet* (London 1996).
- Manning, Patrick, *Big Data in History: A World Historical Archive* (London 2013).
- Mawdsley, E. and T. Munck, *Computing for Historians: An Introductory Guide* (Manchester 1993).
- Middleton, R. and P. Wardley, ‘Information technology in economic and social history: the computer as philosopher’s stone or Pandora’s box?’, *Economic History Review*, 43 (4), (1990), pp. 667–696.
- Weatherill, L. and V. Hemingway, *Using and Designing Databases for Academic Work. A Practical Guide* (Newcastle 1994).

For examples of current computer use in historical research see the journals *International Journal of Humanities and Arts Computing* and *Historical Methods*.

Exercises for Chapters 8 and 9

These exercises are different in nature to those for earlier chapters because they involve relating the statistical techniques covered in this book to econometric modelling exercises and to other computer applications. Accordingly, what follows is a list of articles chosen to represent different recent econometric approaches and applications of ICT to historical research. The exercises are intended to assist academics teaching the subject who might wish to assess their student's capabilities and knowledge, by marking the results. Alternatively, students may wish to discuss their answers within cooperative learning groups.

The list of articles has been carefully researched and chosen to provide models for different sorts of quantitative research applications. We suggest, for an exercise, that you select one or two articles that relate to your own research period or interest and for each write a short assessment (1,000 words) or a fuller review essay (2,000 words).

The questions to be asked in every case are the following:

1. What are the research questions being posed?
2. What sources are being used and how amenable are they to the statistical and other techniques being employed?
3. What pitfalls are there in using the machine-readable data employed here?
4. What statistical, modelling and/or other computing techniques are being used?
5. What difference has the employment of ICT made in the particular piece of research that could not have been accomplished without?
6. Overall, how successful is the piece of research in addressing the research questions posed and in advancing knowledge?

Articles

Bakker, Gerben, 'How motion pictures industrialized entertainment', *Journal of Economic History*, 72 (4), (2012), pp. 1036–1063.

Bateman, Victoria N., 'The evolution of markets in early modern Europe, 1350–1800: a study of wheat prices', *Economic History Review*, 64 (2), (2011), pp. 447–471.

Bignon, Vincent, Rui Esteves and Alfonso Herranz-Loncan, 'Big push or big grab? Railways, government activism, and export growth in Latin America, 1865–1913', *Economic History Review*, 68 (4), (2015), pp. 1277–1305.

Bogart, Dan, 'Turnpike trusts and property income: new evidence on the effects and legislation in eighteenth-century England', *Economic History Review*, 62 (1), (2009), pp. 128–152.

Bonneuil, Noël, Bringé Arnaud and Paul-André Rosental, 'Familial components of first migrations after marriage in nineteenth century France', *Social History*, 33 (1), (2008), pp. 36–59.

- Cantoni, Davide and Noam Yuchtman, 'Medieval universities, legal institutions, and the commercial revolution', *Quarterly Journal of Economics*, 129 (2), (2014), pp. 823–887.
- Chaudhary, Latika and Manuj Garg, 'Does history matter? Colonial education investments in India', *Economic History Review*, 68 (3), (2015), pp. 937–961.
- Crafts, Nicholas and Nikolaus Wolf, 'The location of the UK cotton textiles industry in 1838: a quantitative analysis', *Journal of Economic History*, 74 (4), (2014), pp. 1103–1139.
- Cunningham, Niall, 'The doctrine of vicarious punishment: space, religion and the Belfast troubles of 1920–22', *Journal of Historical Geography*, 40, (2013), pp. 52–66.
- Dunsford, H. M. and S. J. Harris, 'Colonization of the wasteland of County Durham, 1100–1400', *Economic History Review*, 56 (1), (2003), pp. 34–56.
- Graddy, Kathryn, 'Taste endure! The rankings of Roger de Piles (d. 1709) and three centuries of art prices', *Journal of Economic History*, 73 (3), (2013), pp. 766–791.
- Hearn, Brian, Jörg Baten and Dorothee Crayen, 'Quantifying quantitative literacy: age heaping and the history of human capital', *Journal of Economic History*, 69 (3), (2009), pp. 783–808.
- Huff, W. Gregg, 'Boom-or-bust commodities and industrialization in pre-World War II Malaya', *Journal of Economic History*, 62 (4), (2002), pp. 1074–1115.
- Lahey, Joanna N., 'Birthing a nation: the effect of fertility control access on the nineteenth century demographic transition in America', *Journal of Economic History*, 74 (2), (2014), pp. 482–508.
- Lehmann, Sibylle H., 'The German elections in the 1870s: why Germany turned from liberalism to protectionism', *Journal of Economic History*, 70 (1), (2010), pp. 146–178.
- Levin, Naomi, Ruth Kark and Emire Galilee, 'Maps and the settlement of southern Palestine, 1799–1948: an historical/GIS analysis', *Journal of Historical Geography*, 36 (1), (2010), pp. 1–18.
- McLeman, Robert, Sam Herold, Zoran Reljic and Daniel McKenney, 'GIS-based modelling of drought and historical population change on the Canadian prairies', *Journal of Historical Geography*, 36 (1), (2010), pp. 43–56.
- Nuvolari, Allesandro and Michelangelo Vasta, 'Independent invention in Italy during the liberal age, 1861–1913', *Economic History Review*, 68 (3), (2015), pp. 858–886.
- Scott, Peter M. and James Walker, 'Working class household consumption smoothing in interwar Britain', *Journal of Economic History*, 72 (3), (2012), pp. 797–825.
- Voigtländer, Nico and Hans-Joachim Voth, 'The three horsemen of riches: plague, war, and urbanization in early modern Europe', *Review of Economic Studies*, 80 (2), (2012), pp. 774–811.

GLOSSARY

This glossary is fully revised since the first edition with many new terms added including some that readers might encounter in secondary research but that are not directly covered in this volume. In these cases readers are referred to the more advanced statistics handbooks in the reading lists to Chapters 5 and 6, and especially to Charles Feinstein and Mark Thomas, *Making History Count: A Primer of Quantitative Methods for Historians* (Cambridge 2002).

age heaping: generally occurs in demographic and other historical data such as censuses where a preference is shown for stating ages not accurately but by rounding up or down to a figure ending in 0 or 5. In some cultures or circumstances heaping might occur because of a preference for ages that mean something favourable, that avoid certain taxes or increase certain benefits, for example. Where demographic data is suspected of age heaping around 0 or 5 the degree of age heaping can be measured and is sometimes used as a proxy for innumeracy levels. Where age heaping is suspected but the historian wishes to correct the data to eliminate the distortion the **Whipple index** can be used.

age pyramid: figure which represents the age structure of males and females in a population by placing histograms of each back to back, thus allowing immediate visual comparison.

agency theory: a theory that encapsulates the relationship between principals and agents in business and is concerned with identifying problems that can exist in agency relationships, that is, between principals and agents charged with conducting business on account of the principal.

aggregate data: information about a group which produces, or is in the form of, a total.

algorithm: a step-by-step procedure or formula for solving a problem.

anthropometric studies: studies that rely on measurements of biological characteristics such as heights and weights, usually as a proxy for other more important variables, for example relying on time series of heights to consider changes in nutrition and living standards.

arithmetic mean: see **mean**.

asymmetric information: a situation in which one party in a transaction or relationship has more or superior information than the other or others. This often happens in transactions where the seller knows more than the buyer, although the reverse can also be the case.

autocorrelation: a distortion introduced into regression and correlation results with time series data because of the impact of non-random errors.

average: a measure of the central tendency of a distribution, usually calculated as the arithmetic mean, the median or the mode.

back-projection: a technique used to infer or estimate figures for periods where data is unavailable, on the basis of later periods where the figures do exist.

bar chart: a diagrammatic method for displaying frequency distributions in which bars of equal width are drawn to represent each category, with the length of each bar being proportional to the number, or frequency of occurrence, of each category.

base period (usually base year): the case in a time series used as the point of view in looking at other cases, and the yardstick against which all other cases in an index are measured.

'basket of goods': a list of the components of expenditure of an average individual or family which is used to form a measure of the cost of living and changes in it.

bias: any situation in which the accuracy, reliability and validity of historical data or research results are distorted by the limitations of a research method or by a researcher's predispositions.

Glossary

In a narrower statistical sense, bias is the difference between a hypothetical ‘true’ value of a variable in a population and that obtained from studying a sample.

Big Data: resources of quantitative (and other) information, stored electronically and often amounting to several terabytes. Big Data often draws comparative evidence from many parts of the world. The volume and diversity of data make it difficult to store, retrieve, analyse and utilize.

bio-metrics: the statistical study of biological phenomena.

Boolean operators: computer retrievals based on the operations retrieve all ‘X’ AND ‘Y’ or retrieve all ‘X’ NOT ‘Y’, named after the nineteenth-century logician George Boole.

Borda ranking/Borda count: ranking and counting each candidate or alternative with 1 point for each last placing received, 2 points for each next-to-last placing, and so on etc., all the way to N points for each first-place vote (where N is the number of candidates/alternatives).

Breusch–Gordon Lagrange Multiplier Test: a test for autocorrelation which is used where autoregression is suspected. In this case the Durbin–Watson test would be inappropriate and misleading. The Breusch–Gordon Lagrange Multiplier Test is not covered in this volume and readers are directed to other sources such as Feinstein and Thomas, *Making History Count*, pp. 316, 448.

cartogram: a map on which pie charts, graphs or other symbols are superimposed to represent quantities or variables.

case: a unit of study around which information is gathered and arranged.

categorical data: gives qualitative information only, though this may be ranked in some kind of hierarchy, for example status or value.

cell: a box containing numbers or text in a data matrix or table: the intersection of a row and a column.

central tendency: measures of central tendency are those that seek the most typical experience or characteristic in a dataset. With numerical data, the measure of central tendency will normally be the arithmetic average or mean, the median (the middle point when ranked in size order), or the mode (the most commonly occurring characteristic or measure).

ceteris paribus: a term meaning ‘other things being equal’. It is much used by economists and econometric historians interested in examining the importance of one amongst several independent variables which are acting simultaneously in a model. Statistical techniques enable one to simulate the impact of just one independent variable by holding constant the effect of the other variables.

chart: a common word for graph (see **graph**).

chi-square: a statistic that measures the distribution of the divergence between observed and expected results in a frequency table.

class interval: the range of values of a category of information in a grouped frequency distribution.

cliometrics: the quantitative study of history, often used synonymously with econometric history.

cluster sample: a sample selected in a non-random manner because of practicalities or ease of access to such cases.

Cobb–Douglas production function: a particular form of the production function, widely used to represent the technological relationship between the amounts of two or more inputs, particularly physical capital and labour, and the amount of output that can be produced by those inputs.

Cochrane–Orcutt Correction: a correction applied to data where autocorrelation is suspected. It is not covered in this volume and readers are referred to Feinstein and Thomas, *Making History Count*, pp. 315, 373.

coding: a standardized abbreviation assigned to a piece of information, for example F for Female, M for Male.

coefficient of determination: represents the degree to which the movement of one variable is associated with variation in another.

- coefficient of variation:** a measure of the extent to which a variable differs from its mean. It is the standard deviation of the distribution expressed as a percentage of the mean.
- composite index:** a series formed by blending several component series together in a weighted combination. See also **weights, index/indices**.
- concordance:** a record of the principal words used in a book or body of work, listing every instance of each word with its immediate context. There are various software packages that enable this process to be undertaken.
- constant prices:** a valuation time series which has been adjusted to allow for the effects of inflation or deflation. Normally this is done by using the prices obtained in a given base year rather than those current for each time period.
- contingency coefficient:** a measure of relationship between two variables which have been tabulated in a **contingency table**.
- contingency table:** a table in which two variables are plotted against one another or cross-tabulated.
- continuous data:** interval data, such as heights, weights or wages, measured on a scale which includes fractions or decimals as well as whole numbers.
- correlation:** the association between two variables such that when one changes in magnitude, the other does also: there is a concomitant variation.
- correlation coefficient:** a measure of the association between two variables. The nearer the correlation coefficient is to 1, the greater is the strength of the relationship between the two variables.
- cost of living index:** a measure of the movement of prices of a collection of consumer goods regarded as typically demanded by households.
- cost-benefit analysis:** a comparison of all the important costs and benefits (including social costs and benefits expressed in monetary valuations) of the innovation of a particular institution or technology over a stated time period or at a certain date.
- counterfactual:** a hypothetical event or state of being which is measured or described for the purposes of evaluating the costs or benefits of what really happened against what might have happened in the 'second best' scenario.
- counterfactual history:** the calculation of costs and benefits of a particular innovation event or institution in the past compared with the costs and benefits that would have obtained in its absence and in the light of alternatives.
- covariance:** a measure of how closely two variables move together with no necessary implication that the two may be related.
- cross-sectional data:** information on cases that are not in a time series: refers to measures at a particular point in time.
- cumulative frequency distribution:** a **frequency distribution** in which the sum of frequencies in a specific category or class together with the frequency in all categories or classes below it is given.
- cycle:** a boom and slump (growth followed by decline) in a time series. Cycles, often with an identifiable periodicity, are often repeated many times throughout a time series.
- cyclical:** exhibiting a regular movement.
- data:** information relating to cases.
- data matrix:** a way of organizing a dataset in the form of a table with rows and columns.
- data processing:** producing results using a computer.
- database:** a collection of related data, organized in a predetermined manner, according to a set of logical rules. Normally the data is arranged in tabular form, which contains discrete categories of information called fields for a number of distinct and unique cases. A database can contain one or more tables and is normally stored in a computer.
- database management system (DBMS):** a computer application which allows the user to create, manage and analyse electronic tables in a database.

Glossary

- dataset:** information relating to cases that the researcher selects in order to address a particular question.
- decile:** one of nine actual or notional values of a variable dividing its distribution into ten groups with equal frequencies.
- deduction:** the process of building up knowledge by testing theories or ideas against the facts (cf. **induction**).
- denominator:** the lower element in a fraction.
- dependent variable:** a variable, the measure of which is determined by the movement of another variable or other variables under consideration.
- descriptive statistics:** statistics concerned with summarizing or describing a distribution or a sample. They consist of methods of statistical display and rearrangement which contribute to clarity of information and often provide a basis for initial analysis of the figures.
- detrended series:** a time series with any underlying long-term rise or fall over time removed to reveal more clearly regular and irregular fluctuations.
- dichotomous variable:** a variable which can only take two values, for example, sex which takes either male or female.
- discrete data:** interval data which can be expressed only in whole numbers, for example, numbers of people.
- distribution:** a range of values observed for any one variable.
- dummy variables:** are used to recode some sorts of categorical data (most commonly dichotomous variables) in numerical form.
- econometric history:** application of economic theory and the methods of mathematical statistics in economic history.
- econometrics:** the application of mathematical statistics to economics.
- elementary descriptive statistics:** various techniques of statistics that are concerned to identify and demonstrate the most important features of any given data. They are largely concerned with rearrangement, and tabular or graphical display.
- endogenous:** an endogenous factor or variable is one which is generated and acts from within the system or model under investigation. For example, population growth or industrial production within an economy, or the impact of real wage levels or levels of trade union organization upon the size of public demonstrations.
- European Economic Community (EEC):** a free trade area with a common external tariff. There were six original member states, which increased to nine (including the UK) in 1973, to 12 in the 1980s and to 15 in the 1990s.
- exogenous:** an exogenous factor or variable is one which acts from outside of the particular system or model being examined. For example, the impact of external trade on an economy or the impact of weather upon the size and nature of public demonstrations or soccer crowds.
- expert systems:** advanced softwares that employ artificial intelligence to solve specialized problems.
- exponential scale:** a vertical scale on a graph where units are successively the square roots of the previous unit.
- externality:** an economic or social cost or benefit which is not generally included in the market or price evaluation of an innovation or process. For example, increased disease and death rates experienced by migrants to towns and cities in the process of industrialization; the noise, traffic and environmental pollution resulting from the building of additional airport runways; the wider social benefits of public sector broadcasting.
- extrapolation:** the estimation of missing values of a variable (on a regression line, for example) based on the trend apparent in the known values.
- field:** information contained in one column of a database table.
- filter** (in a database): applies a set of criteria to show a subset of the records or to sort the records, for example, in a parish register database to show all baptisms of legitimate children before 1730 (by filtering out records marked as illegitimate or after 1730).

Fisher's Ideal Price Index: is the geometric mean of the Paasche and Laspeyres indices. It uses both current year and base year quantities as weights. The index corrects the positive bias inherent in the Laspeyres index and the negative bias inherent in the Paasche index.

flatfile (database): a simple database consisting of just one table.

fluctuation: a marked, often cyclical, movement of a variable in a time series.

frequency: the number of times that a particular value of a variable appears.

frequency curve: is formed from a frequency polygon by using much smaller class intervals so that the line of the graph is smoothed into a curve.

frequency distribution: the number of times each value of a variable occurs in a set of observations. Often the observations are arranged in groups and the frequency distribution is displayed in a table.

frequency polygon: another way to show the information in a frequency table. It is derived graphically from a histogram and is formed by plotting values at the midpoints of each class, joining these together with straight lines.

geographic information system (GIS): a system designed to capture, store, manipulate, analyse, manage, and present many types of spatial or geographical data.

geometric mean: the N th root of the product of a distribution (where N is the number of items in the distribution).

Gini coefficient: a summary measure of distributional inequality. It is a measure of statistical dispersion intended to represent the income distribution of a nation's residents, and is the most commonly used measure of inequality.

graph: a figure which relates the movement of one variable to another or charts the movement of one or more variables over time.

growth rate: measure of the speed of expansion in an economy, an industry or a sector. The measure may be negative as well as positive.

hedonic prices: prices based on the principle that, the price of a marketed good is affected by certain external environmental or perceptual factors that can raise or lower the price of that good. This is commonly applied to the housing market because the price of a house can be affected by factors such as scenic views, house appearance and neighbourhood demand. In modern GDP per capita time series calculations, hedonic prices are often used to reflect the real value rather than the prevailing price of important innovative products and services.

histogram: a diagrammatic representation of a frequency distribution consisting of a series of rectangles or bars with a width proportional to the class interval concerned and an area proportional to the frequency.

hypothesis: a working theory relating to cause, effect or change.

hypothetical: made-up, unreal but useful as a reflection of reality.

independent random sample: a sample selected in such a way that each case has an equal chance of being chosen.

independent variable: a variable, the value of which is not determined by the movement of another variable, or other variables, under consideration.

index (pl. indices): a way of recording variation in a time series by converting all values to a percentage of the value in a certain base year/day/month. This is especially useful in highlighting movements in one or more time series when the original units were complex or different from one another. See also **composite index, real index**.

index number problem: the problems of selecting the elements in a composite index and of assigning accurate weights to each variable.

induction: the process of building up knowledge by generalizing on the basis of facts or data which have been assembled (cf. **deduction**).

inferential statistics: statistics concerned with generalizing from a sample, to make estimates and inferences about a wider population.

Glossary

inflation: price increases which reduce the purchasing power of money wages and especially of fixed incomes such as pensions and which make it difficult to sell goods competitively in external markets.

instrumentalism: judging theories on the basis of their predictive ability rather than their relationship to reality in other respects.

Interactive Structured Query Language (ISQL): a series of computer commands that enable the calling up of certain vectors of information relating to particular cases in a table or tables, the ordering and summing of such vectors, and other more complicated instructions involving highly selective cells of information and the linking of one record with another, often across different datasets within the database.

interquartile range: a measure of dispersion around the median: the range of the middle 50 per cent of values of a dataset which is arranged in rank order.

Laspeyres index: a composite index which uses base year quantities and weights based on estimated weights in the base year.

life tables: tables which present estimates of longevity (under various conditions, for occupational groups, and so on), based upon probabilities of death rates calculated from experience.

line of best fit: a line on a graph (often the linear trend in a time series or a regression line in a scatter graph) which is drawn in such a way that all the distances of observations above the line equal all those below.

log linear analysis: a technique of statistical analysis which transforms non-linear models into linear models by the use of logarithms. This is necessary because social data is often nominal or ordinal and therefore does not meet the assumptions needed for many statistical techniques. It is a causal modelling device involving setting up models to test against the data, successively adjusting the model until the best fit is found.

log scale (or logarithmic scale): a non-linear scale used when there is a large range of quantities. The logarithm of a physical quantity is used in a figure or graph instead of the quantity itself.

logarithms: representations of numbers generally expressed as a power or exponent of 10. Their use facilitates multiplication and division calculations for large or complex numbers but they have now been superseded for this purpose by calculators and computers.

logistic growth curve: a pattern of growth common in studies of social phenomena where growth begins slowly, increases rapidly and finally stabilizes.

Logit analysis: a form of regression in time series analysis which assists in avoiding the assumption of uniform linear change.

Logit models: models that have been developed to aid regression analysis where there is a limit upon the values of the dependent variables. This is quite common in time series data. Such models avoid the assumption of uniform linear change which lies behind normal time series regression. In these cases it is wise not to employ standard regression techniques and most computer packages include the facility instead to incorporate logit tools. See also Probit and Tobit models which are alternative substitutes for standard regression techniques and which might be preferred in slightly different circumstances. Such tools are not examined in this volume; instead readers might refer to Feinstein and Thomas, *Making History Count*, Chapter 13.

Lorenz curve: a cumulative percentage curve (usually of the income of a nation). The shape of the curve, its deviation away from a straight line, is a visual indication of income inequality.

macroeconomic: refers to a whole economic system that is the aggregate of the behaviour of individual economic agents and which generally exhibits regularities of character and behaviour different from those identifiable at the individual level.

Mann–Whitney test (or the Mann–Whitney *U* test): a nonparametric test of the null hypothesis that two samples come from the same population against an alternative hypothesis (usually that one population tends to have larger values than the other). It can be applied to unknown

distributions contrary to the *t*-test which can be applied only to normal distributions, and it is nearly as efficient as the *t*-test on normal distributions.

matrix: an arrangement of data where the variables' relation to cases are arranged in columns and the cases arranged in rows to provide cells of information.

mean or arithmetic mean: a measure of average which is found by adding all the values and dividing by the number of cases.

mechanical objectivity: freedom from bias that derives from the accuracy of scientific measurement, an accuracy that can be tested independently and, ideally, where results are capable of being reproduced.

median: a measure of average formed by ranking all observations in size order and taking the middle reading (or the mean of the two middle readings in the case of an even number of observations).

microeconomic: refers to the actions of individual economic agents and their choices as producers and consumers.

modal class: the class containing the most observations (or the highest frequency of observations).

mode: a measure of average: the most frequently occurring observation in a dataset.

model building: in statistical and economics work this term generally refers to attempts to simulate the interactions of various variables in an algebraic equation or series of equations.

moral statistics: numerical data that are held to be indicative of social conditions or social problems. These include, for example, statistics of illiteracy, family size, fertility, diseases and mortality, illegitimacy, abortion, divorce, prostitution and poverty. Such data were and are widely cited in debates about social reform.

moving average: an average for the movement of a variable over several years (or other time periods). It is called moving because as the time periods pass the earlier ones are dropped from the average and the mid-point of the average moves. It is a method of smoothing out fluctuations in time series data so that the longer term trends can be more easily observed.

multicollinearity: a distortion introduced into regression and correlation results with time series because of common trend or cyclical elements in the movement of the variables.

multivariate analysis: statistical procedures involving more than one dependent variable.

negatively skewed: description of a distribution where most observations lie above the mean.

neoclassical economics: the dominant form of theorizing in Western economics since the late nineteenth century. Based upon logical analysis of the rational profit-seeking behaviour of large numbers of well-informed individuals active in markets governed by legal systems which enforce property rights and contracts.

nominal data: categorical, qualitative information where the order in which it appears is not important.

non-parametric statistics: statistical methods for the analysis of ordinal and categorical sample data which do not require assumptions about the shape of the distribution from which the samples have been drawn.

normal distribution or error curve: an ideal-type continuous distribution of a variable. The normal curve is bell-shaped and the mean, median and mode are equal.

normative theory: any theory which seeks to establish the values or norms that best fit the overall needs or requirements of society.

null hypothesis: used in calculations of the degree to which two variables may be related to one another. This hypothesis relates to the distribution of the variables as they would be if there were no relationship at all.

numerator: the upper element in a fraction.

numeric data: data expressed in numbers.

objectivity: (i) accounts of the external world held to represent the world as it exists independently of our conceptions; (ii) a more frequent usage is knowledge claimed to meet criteria of validity and reliability and held to be free from bias: avoiding subjectivity by following impartial rules

Glossary

of measurement, observation and experiment (in the process ignoring what experience, intuition and moral inclination suggest to be correct). This **mechanical objectivity** involves personal restraint and following the rules. To some extent this describes quantification and the language and discipline of quantification.

OCR (optical character recognition): the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-readable form, important in the digitization of historical sources and evidence.

ordinal data: categorical, qualitative data where the order in which it appears is of some importance.

outliers: atypical cases at the extremes of a distribution.

Paasche index: a composite index which uses the current year or series-end quantities and weights, that is, based on estimates of weights at the end of the series.

parametric statistics: a sample of figures where we assume that the parent population has a normal distribution (in reality a normal distribution is only approximated but this is regarded as acceptable to fulfil the criteria for parametric analysis).

percentile: used with rank order distributions, these are values that divide the distribution into 100 parts of equal frequency.

pie chart: a way of displaying the distribution of nominal, ordinal or interval data by drawing a circle divided into segments of the appropriate size to represent each class of data.

political arithmetic: a new way of viewing society and of analysing social and political issues associated first with William Petty from the late seventeenth century. The foundation of political arithmetic was the idea that the prosperity and strength of the state rested on the number and condition of its subjects.

population: the entire group of subjects to which a researcher intends the results of a study to apply; the larger group to which inferences are made on the basis of the particular set of people studied.

positively skewed: description of a distribution where most observations lie below the mean.

Positivism: a doctrine formulated by Auguste Comte which asserts that the only true knowledge is scientific knowledge (the study and explanation of observable phenomena whether natural or social). Positive knowledge of social phenomena was expected to encourage scientifically grounded intervention in social and economic affairs which would transform social life.

probability: a number ranging from 0 (impossible) to 1 (certain) that indicates how likely it is that a specific outcome will occur in the long run.

probability theory: a list of rules for calculating the probabilities of complex events. It predicts how variables are likely to behave and provides a numerical estimate of that prediction. In history and social science probability theory is important in relation to sampling.

Probit models: see **Logit models**.

proportional sample: see **systematic sample**.

proxy: a 'stand-in'. In statistics this refers to the figures that we must use (because they are available) instead of the figures that we would ideally like to use for the analysis at hand but that are not available.

psephology: the scientific/statistical study of past elections and voting. Word introduced by British historian R. B. McCallum in 1952.

p-statistic or p-value: the probability of obtaining a result equal to or 'more extreme' than what was actually observed, assuming that the hypothesis under consideration is true. If the p-value is equal to or smaller than the selected significance level, it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true and thus that hypothesis must be rejected.

pyramid chart: a triangular chart used to illustrate the distribution of a small number of variables in a population where categories have a clear hierarchy and where the proportion of the population in each category varies inversely with rank.

quartile: the points at which a rank ordered series divides into four equal parts.

quartile deviation: the interquartile range of a distribution divided by 2 (this is also termed the semi-interquartile range).

quintile: the points at which a rank ordered series divides into five equal parts.

random sample: see **independent random sample**.

range: the spread of a set of data (the highest value of the distribution minus the lowest).

rank order: the distribution of a variable ranked in size order from the smallest to the largest.

rationality: a culturally constructed set of concepts and ideas about moral behaviour. In neoclassical economic theory the rationality postulate states that if an individual is presented with a situation of choice in an economic setting he or she will act to optimize his or her economic position.

real index: an index which has been adjusted (deflated or inflated) to allow for the movement of another series, most often prices.

real movements: a money value adjusted for changes in prices. Used commonly in economics where a variable, for example, real wages, has been adjusted for inflation or deflation of prices (so that the purchasing power of money wages is in this case being expressed). To convert money values to real values requires the use of an appropriate deflator, for example the Retail Price Index which is the official index of change in consumer prices in Britain.

regression: a technique for analysing the relationship between two or more interval level variables in order to predict the value of one from variation in the other(s).

regression coefficient (also known as the slope coefficient): the slope of the regression line. It is the coefficient by which the dependent variable moves in response to the independent variable.

regression line: a line which represents the closeness and pattern of movement between two variables. It can be drawn as the least squares line of best fit through a scatter graph. It represents the best estimate of the relationship between the variables based upon the available evidence.

relational (database): a database involving many tables with the possibility of linking data across files and interrogating several tables simultaneously.

relational database management system (RDBMS): a program that allows one to associate data in two or more tables on the basis of common fields or cases.

sample: a selection of data from the whole population.

sampling: the procedures used to extract a number of cases for analysis from a larger population.

The general aim is to generate a sample that is representative of the population as a whole.

sampling error: the difference between the 'true' value of a characteristic within a population and the value estimated from a sample of that population.

sampling theory: statistical theory which enables estimation of the degree to which sample results reflect or vary away from results which would have been obtained had the whole population been examined.

scatter graph: a graph in which pairs of variables are plotted as an initial indication of the extent to which there is a relationship between the two.

seasonal: description of variations in data which occur because of changing climatic or other factors which regularly change over the course of a year.

seasonality: a regular seasonal pattern of fluctuations that may be present in a time series.

semi-interquartile range: see **quartile deviation**.

semi-logarithmic graph: a graph that has a logarithmic scale on the y -axis, and a linear scale on the x -axis.

serial history: historical research based upon the study of long-term movements in vital variables such as agricultural output, population indices, prices, wages.

series: a run of data relating to one variable.

series of first differences: used to reduce the impact of a common trend when comparing two or more series. The various series are often converted to series of first differences formed by subtracting each value from its predecessor.

Glossary

significance: a much-debated and highly problematic term in statistics which derives from the degree of probability or chance associated with an occurrence and which can be expressed at various levels. The real significance of statistical occurrences, particularly relationships between variables, can, however, only be assessed by experts for whom statistics is a tool or guide rather than an answer.

significance test: a test of the probability of an observed result occurring by chance. The result of the test is expressed as a statistic (for example, *t*-ratio, *p*-statistic) which can be assessed against different levels of probability.

skewed distribution: a distribution where observations are distributed very unevenly around the mean.

slope coefficient: see **regression coefficient**.

social savings: the benefit from a project or innovation to society compared with an alternative or second-best development.

spreadsheet: software for organizing and processing numerical information arranged in a matrix table of rows and columns.

SQL (Structured Query Language): the most common query language encountered in DBMS. Developed by IBM, it consists of basic commands such as INSERT, SELECT, UPDATE and DELETE, which can be enhanced by using an array of supplementary commands.

standard deviation: a measure of dispersion of a distribution around the mean, normally represented by the letter *s*.

statistics: before the mid-nineteenth century, statistics was an ill-defined ‘science’ of states and conditions; of data (numerical and other) relating to the wealth and power of the state. Standard usage by the 1830s and 1840s was that statistics referred to numbers of things and it came gradually to become an empirical, usually quantitative, science. The term came to be applied to a field of applied mathematics only in the twentieth century. Today it is concerned with scientific methods for collecting, organizing, summarizing and presenting quantitative data.

stratified sample: a sample which is deliberately selected proportionally to represent the different classes or categories of the wider population.

structured data: a source of historical information which is organized into clearly defined categories.

Substantivism: a set of ideas associated originally with Karl Polanyi which places stress upon the need to understand economies and economic exchanges in their own terms and not through the lens of our own time and culture.

survey: information collected by interviews or questionnaires with respondents chosen in a variety of possible ways but subject to the constraints (in terms of their typicality) formed by willingness to participate and availability.

surviving sample: data which is chosen for analysis because the records have survived but not because they necessarily reflect the experience of the whole population of cases.

systematic sample: a selected proportion of the population, chosen in such a way as to avoid bias as much as is possible.

text analysis software: software that enables users to determine the frequency with which words or phrases are used, create concordances, view words in context, and otherwise study patterns in texts.

textual data: information in non-numeric form, that is, letters, words, prose.

time series: a dataset which comprises the movement of a variable or variables over time.

Tobit models: see **Logit models**.

trend: a straight line that best expresses the direction of movement of a time series over the longer term.

trend line: a line of best fit through a time series positioned in such a way that all the distances of observations above the line equal all those below.

- T-statistic:** a ratio of the departure of an estimated parameter from its notional value and its standard error. It is used in hypothesis testing, for example in the student's **t-test**.
- t-test:** a hypothesis testing procedure in which the population variance is unknown; it compares *t* scores from a sample to a comparison distribution called a *t*-distribution to give levels of statistical significance for sample results.
- variable:** a characteristic which can be measured. It may vary along a continuum (**continuous variable**) as with heights or weights; be discrete, that is, measured only in single units (**discrete variable**) as with household size; or be **dichotomous** as with sex.
- variance:** a measure of the dispersal of a distribution which is formed from the average of the squares of the deviations of observations from the mean.
- Variance Inflation Factor (VIF):** quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.
- vector:** a column of numerical information, normally expressing the movement of a variable relating to a number of different cases.
- views:** screens of information that can be called up from a database.
- virtual history:** hypothetical history derived from assessing and imagining what might have been the result if particular historical events, crises or conjunctures had had an outcome very different from that which was obtained in reality.
- weights:** measures reflecting the relative importance of an item in a composite index.
- Whipple index** (also known as the index of concentrations): a method that measures the tendency for people to state their age inaccurately in a historical record. Mostly the concern is with the tendency to round up or down to a number ending in 0 or 5. The index score in this case is obtained by summing the number of persons in the age range 23 to 62 inclusive, who report ages ending in 0 and 5, dividing that sum by the total population between ages 23 and 62 years inclusive, and multiplying the result by 5. Restated as a percentage, index scores range between 100 (no preference for ages ending in 0 and 5) and 500 (all people reporting ages ending in 0 and 5).
- word cloud:** an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.
- x-axis:** the horizontal axis on a graph (which normally takes the independent variable, for example time which is not affected by other variables).
- y-axis:** the vertical axis on a graph (which normally takes the dependent variable(s), i.e. those that may be affected by other variables).
- Z score:** the number of standard deviations an observation is above (or below, if it is negative) the mean in its distribution.

NOTES

Preface

1. Roger Middleton, *The British Economy Since 1945: Engaging with the Debate* (London 2000), Appendix III.
2. For illustrations of this disenfranchisement see John Allen Paulos, *Innumeracy. Mathematical Illiteracy and its Consequences* (London 1988).
3. One exception among early social science statistics textbooks and still in print is Francis Clegg, *Simple Statistics: A Course Book for Social Sciences* (Cambridge 1983), complete with amusing cartoons. More recently, Sonja Cameron and Sarah Richardson, *Using Computers in History* (Basingstoke 2005) provides a short and simple introduction to some basic statistical processing alongside the use of computers for presenting historical research, to accessing information on the Internet and to acquiring and manipulating digital images. Of the bewildering variety of more involved social science statistics texts the most straightforward and recent is Liam Foster, Ian Diamond and Julie Jeffries, *Beginning Statistics: An Introduction for Social Scientists* (2nd edition London 2014). Catherine Marsh, *Exploring Data. An Introduction to Data Analysis for Social Scientists* (Cambridge 1988 and reprints) remains a useful social science text incorporating much discussion of British data. Texts introducing quantitative methods to historians since the 1990s have generally been more technically advanced than most historians require. The best is Charles Feinstein and Mark Thomas, *Making History Count. A Primer in Quantitative Methods for Historians* (Cambridge 2002), which is linked to electronically available data and examples on the publisher's website; and L. Haskins and K. Jeffreys, *Understanding Quantitative History* (Cambridge, MA 1991, 2011 reprint). The latter, oriented towards North American data and needs, is more advanced than the present volume but worth consulting. The structure of *History by Numbers* and the level at which it is pitched owes much to Roderick Floud, *An Introduction to Quantitative Methods for Historians* (London 1973, 2nd edition 1979), now out of print.
4. Pat Hudson, 'Numbers and words: quantitative methods for scholars of texts' in Gabriele Griffin (ed.), *Research Methods for English Studies* (Edinburgh 2005, 2nd edn. 2013) pp. 131–156.
5. Darrell Huff, *How to Lie with Statistics* (London 1973) p. 10.
6. See Chapter 8.
7. Hudson (2005) 'Numbers and words'.
8. See www.esrc.ac.uk/_images/Undergraduate_quantitative_research_methods_tcm8-2722.pdf consulted 13 September 2015. See also: *National strategy for building a world class social science research base in quantitative methods* ESRC, HEFCE, HEFCW (2010): www.esrc.ac.uk/funding-and-guidance/tools-and-resources/research-resources/initiatives/qmi.aspx (accessed 2 September 2015).
For the ESRC's Quantitative Methods Initiative website detailing learning resources, datasets and what is going on in social science and ICT research, blogs and so on: www.quantitativemethods.ac.uk/ (accessed 12 August 2015).
9. See www.heacademy.ac.uk/node/3470 (accessed 12 August 2015).

Notes

10. British Academy, *Society Counts. Quantitative Skills in the Social Science and Humanities* (London 2012).
11. See [www.nuffieldfoundation.org/sites/default/files/files/Learned%20society%20supporting%20statement%20on%20QS%20\(1\).pdf](http://www.nuffieldfoundation.org/sites/default/files/files/Learned%20society%20supporting%20statement%20on%20QS%20(1).pdf) (accessed 2 July 2015).
12. See www.nuffieldfoundation.org/sites/default/files/files/QM%20Programme%20Background_v_FINAL; www.hefce.ac.uk/kess/qss/ (accessed 12 August 2015).
13. See the online appendix at www.bloomsbury.com/history-by-numbers for an indication of what is currently available. There are of course dangers with using all this easily available data, collected and stored with different levels of accuracy, professionalism and compatibility: see for example the critique posed by Tim Hitchcock in 'Confronting the digital or how academic history writing lost the plot', *Cultural and Social History*, 10 (1), (2013), pp. 9–23.
14. For some recent examples and discussion of the digital turn in early modern cultural history see the special issue of the *Journal of Early Modern Cultural Studies*, 13 (4), (2013).

Chapter 1 The Prospects and Pitfalls of History by Numbers

1. For an excellent collection of essays on the power and the problematic nature of statistics generated to record, analyse or aid in the improvement of various aspects of economy and society over time see Tom Crook and Glen O'Hara (eds), *Statistics and The Public Sphere: Numbers and the People in Modern Britain, 1800–2000* (London 2011).
2. Some common statistical techniques are mentioned in this chapter. They will all be more fully explained later in the book. If difficulties are experienced in understanding the arguments made here, use of the Glossary is recommended.
3. Pat Hudson, 'Numbers and words: quantitative methods for scholars of texts', in Gabriele Griffin (ed.), *Research Methods for English Studies* (Edinburgh 2005), pp. 131–156.
4. François Furet, 'Quantitative history', *Daedalus*, 100, quoted by Carlo Ginzburg, *The Cheese and the Worms* (London 1971), p. xx.
5. E. Le Roy Ladurie, *The Peasants of Languedoc* (English translation, London 1974). This is only one of many Annales studies of different French regions using *histoire sérielle*.
6. For a contemporary survey and collection of examples see Peter Temin, *The New Economic History: Selected Readings* (London 1973).
7. See Chapter 4.
8. Lawrence Stone, *The Past and the Present Revisited* (London 1987), p. 94. For a longer sustained attack on the quantifiers see J. Barzun, *Clio and the Doctors* (Chicago 1974).
9. These attempts have come from both sides. Economic sociologists have developed some very interesting sophistications of 'economic' theory most of which have been virtually ignored by economists. See, for example, N. J. Smelser and R. Swedberg, *Handbook of Economic Sociology* (New York 1994). A significant grouping of economists, on the other hand, have reached out to 'colonize' areas of social and cultural history in the last few decades investigating many topics from family relationships to crime and drugs through the prism of market and rational choice theory. See, for example, M. Tommasi and K. Lerulli, *The New Economics of Human Behaviour* (Cambridge 1995).
10. Carl Bridenbaugh, 'The great mutation', *American Historical Review*, 68, (1963), p. 326, quoted by John Tosh in *The Pursuit of History*, 2nd edition (London 1991), p. 202.

11. Tony Judt, 'A clown in regal purple: social history and the historians', *History Workshop Journal*, 7, (1979), p. 74.
12. The phrase is from L. Stone, *The Past and the Present Revisited* (London 1987), p. 94, who also quotes Liam Hudson.
13. R. Cobb, 'Historians in white coats', *Times Literary Supplement*, 3 December 1971.
14. Cobb, 'Historians in white coats', quoted in C. Tilly, *As Sociology Meets History* (New York 1981), p. 72.
15. Tilly (1981) *As Sociology Meets History*, p. 82.
16. Tilly (1981) *As Sociology Meets History*, p. 53.
17. This is relayed in more detail in Chapter 8.
18. For further discussion of this see Chapter 9. For an early statement of possibilities and pitfalls see R. Middleton and P. Wardley, 'Information technology in economic and social history: the computer as philosopher's stone or Pandora's box?', *Economic History Review*, 43 (4), (1990), pp. 667–696.
19. Compare L. Stone, *The Family, Sex and Marriage in England, 1500–1800* (London 1977), a study of elites using diary evidence, with E. A. Wrigley and R. S. Schofield, *The Population History of England, 1541–1871* (Cambridge 1981), which derived statistics from more than 400 parish registers.
20. See for example Julian Hoppitt, *Risk and Failure in English Business, 1700–1800* (Cambridge 1987); R. Lloyd Jones and M. J. Lewis, *Manchester and the Age of the Factory: The Business Structure of Cottonopolis in the Industrial Revolution* (London 1988).
21. Michael Anderson, *Family Structure in Nineteenth Century Lancashire* (London 1971).
22. See www.nappdata.org/napp/resources/publications_pdf/1881gbsample.pdf (accessed 14 August 2015).
23. H-J. Voth and T. Leunig, 'Did smallpox reduce height? Stature and the standard of living in London, 1770–1873', *Economic History Review*, 49 (3), (1996), pp. 541–560; D. Oxley, "The seat of death and terror": urbanization, stunting, and smallpox', *Economic History Review*, 56 (4), (2003), pp. 623–656; T. Leunig and H-J. Voth, 'Comment on "Seat of death and terror"', *Economic History Review*, 59 (3), (2006), pp. 607–616; D. Oxley, 'Pitted but not pitied, or, does smallpox make you small?', *Economic History Review*, 59 (3), (2006), pp. 617–635. The latter article is included as an exercise in this volume (p. 236).
24. The Cambridge Group followed up the 1981 volume analysing aggregate demographic indices, with a collection of their reconstitutions studies: E. A. Wrigley, R. Davies, J. Oeppen and R. Schofield (eds), *English Population History from Family Reconstitution, 1580–1837* (Cambridge 1997).
25. One of the best of these is S. Szczerter, *Fertility, Class and Gender in Britain, 1860–1940* (Cambridge 1996).
26. See www.chia.pitt.edu (accessed 3 September 2015).
27. *The Journal of Economic History* and *Explorations in Economic History* have a very high proportion of articles that employ models alongside quantification.
28. For an introduction to such history see D. N. McCloskey, *Econometric History* (London 1987). For extended but equally accessible discussion see T. G. Rawski (ed.), *Economics and the Historian* (London 1996); for a representative collection of essays using such techniques in the late 1980s see N. F. R. Crafts, N. Dimsdale and S. Engerman (eds), *Quantitative Economic History* (London 1991).

Notes

29. They have been discussed in this way, as part of the rhetoric of economics alongside other devices such as storytelling and metaphor, by D. N. McCloskey in *If You're So Smart; The Narrative of Economic Expertise* (Chicago 1990).
30. For discussion of the pitfalls of official values in measuring foreign trade see G. N. Clark, *Guide to English Commercial Statistics, 1696–1782* (London 1938).
31. For further discussion of these problems with parish register data see Wrigley, Davies, Oeppen and Schofield (1997) *English Population History from Family Reconstitution*.
32. For recent discussion of the problems of the census and other documents in recording women's work see: www.campop.geog.cam.ac.uk/research/projects/occupations/women/ (accessed August 2015).
33. B. R. Wilson, *Religion in Secular Society* (London 1966); M. Vovelle, *Piete baroque et dechristianisation en Provence au 18e siècle* (Paris 1973), both quoted by Peter Burke, *Sociology and History* (London 1980), p. 40.
34. D. C. Coleman, 'History, economic history and the numbers game', *Historical Journal*, 38 (3), (1995), p. 641.
35. This was pointed out by Peter Lindert in 'English living standards, population growth and Wrigley and Schofield', *Explorations in Economic History*, 20 (2), (1983), pp. 131–155.
36. T. M. Porter, 'Making things quantitative', *Science in Context*, 7 (3), (1994), p. 401.
37. The use of anthropological approaches by cultural historians in their efforts to avoid ethnocentric and anachronistic bias has a parallel here with the problems faced by those collecting quantitative data. The problem of increasing the distortion the closer one tries to observe and to measure has been much discussed and is generally referred to as the Heisenberg uncertainty principle (derived from Werner Heisenberg's studies of quantum physics), which stresses the errors inherent in the sensitive scientific measuring implements required for close observation.
38. Theodore M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton 1995), p. 35.
39. Some of the problems with historical crime statistics for England and Wales are discussed in R. M. Morris, "Lies damn lies and criminal statistics": reinterpreting the criminal statistics in England and Wales', *Crime, History and Societies*, V (1), (2001), pp. 111–127; Howard Taylor, 'Rationing crime: the political economy of criminal statistics since the 1850s', *Economic History Review*, 51 (3), (1998), pp. 569–590.
40. Clifford Geertz, *The Interpretation of Cultures* (New York 1973), p. 9.
41. This is pointed out by N. F. R. Crafts in, *British Economic Growth During the Industrial Revolution* (Oxford 1985).
42. For a classic treatment of the potential misuses of statistics see D. Huff, *How to Lie with Statistics* (London 1954).
43. Stephen J. Gould, *The Mismeasure of Man* (London 1981).
44. For further discussion of the pitfalls of statistical significance see Deirdre McCloskey and Stephen Ziliak, for example at www.deirdremccloskey.com/docs/jsm.pdf (accessed August 2015) and in *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice and Lives* (Ann Arbor, MI 2007).
45. The need for a falsifiable hypothesis was stressed by Karl Popper as a way of defending the hypothetico-deductive approach to data and as a way of seeking a mechanical objectivity in scientific enquiry, as free from bias and subjective manipulation as possible. Karl Popper, *Conjectures and Refutations: The Growth Of Scientific Knowledge* (London 1963).
46. Peter Burke (ed.), *New Perspectives on Historical Writing* (Cambridge 1991), p. 15.

Chapter 2 The Origins and Nature of Quantitative Thinking

1. National variations in the pace and degree of development of statistical concerns in relation to practical problems of state, of economic and social reform, national intellectual cultures, and the development of different academic subjects from physics and astronomy to eugenics, geography and economics is a fascinating subject and explored to some degree in T. M. Porter, *The Rise of Statistical Thinking, 1820–1900* (Princeton 1986); idem., *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life* (Princeton 1995) especially Part 3. For the British experience see J. R. N. Stone, *Some British Empiricists in the Social Sciences, 1650–1900* (Cambridge 1997); M. J. Cullen, *The Statistical Movement in Early Victorian Britain: The Foundations of Empirical Social Research* (Sussex 1975); D. A. Mackenzie, *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge* (Edinburgh 1981); E. Higgs, *The Information State in England: The Central Collection of Information on Citizens Since 1500* (Basingstoke 2004); and C. A. Bayley, *Empire and Information: Intelligence Gathering and Social Communication in India, 1780–1870* (Cambridge 1996). See also A. Rusnock, *Vital Accounts: Quantifying Health and Population in Eighteenth Century England and France* (Cambridge 2002); C. Blum, *Strength in Numbers: Population, Reproduction and Power in Eighteenth Century France* (Baltimore 2002); K. H. Roth, *The Nazi Census: Identification and Control in the Third Reich* (Philadelphia 2004); S. Patriarca, *Numbers and Nationhood: Writing Statistics in Nineteenth Century Italy* (Cambridge 1996).
2. This chapter is largely free of statistical terminology. If necessary, use the Glossary and if any parts of the argument of this chapter appear unclear at this point do not worry as they will be more fully explained in subsequent sections of the book which deal with specific techniques of quantitative investigation.
3. Cullen (1975) *Statistical Movement*, pp. 1–16; Higgs (2004) *Information State*.
4. P. Deane, 'Political arithmetic,' in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: A Dictionary of Economics*, 4 volumes (London 1987), pp. 990–993. There were strong authoritarian undertones in this. Petty, for example, proposed that, because the value of an English life could be calculated as far surpassing an Irish life, the wealth of Britain would be augmented by forcibly transporting all Irish men, except a few cowherds, to England. W. Petty, 'A treatise on Ireland' (1687), in C. H. Hull (ed.), *The Economic Writings of Sir William Petty Together with the Observations on the Bills of Mortality More Probably by Captain John Graunt* (Cambridge 1899) volume 2, p. 554.
5. Q. Skinner, 'History and ideology in the English revolution,' *Historical Journal*, 8 (2), (1965), pp. 171, 129; Q. Skinner, 'Thomas Hobbes and his disciples in England and France,' *Comparative Studies in Society and History*, 8 (2), (1965–1966), pp. 153–167; Porter (1986) *Rise of Statistical Thinking*, p. 19.
6. William Petty, 'Political arithmetic,' Preface in C. H. Hull (ed.), *The Economic Writings of Sir William Petty* (2 volumes, Cambridge 1899), 1, p. 244.
7. Whether Graunt or Petty wrote this work is a matter of some dispute: see Cullen (1975) *Statistical Movement*, p. 2.
8. This was the dominant definition of political arithmetic in the eighteenth century. C. Davenant (1771) *The Political and Commercial Works*, volume 1, p. 128 quoted by J. Hoppit, 'Political arithmetic in eighteenth-century England,' *Economic History Review*, 49 (3), (1996), p. 517 (note 9).
9. Hoppit (1996) 'Political arithmetic,' pp. 516–540.
10. This was a protracted business – a struggle between scientific ideals and administrative practicalities: J. Hoppit, 'Reforming Britain's weights and measures, 1660–1824,' *English Historical Review*, 108 (426), (1993), pp. 82–104. It met with considerable opposition, see for example: R.

Notes

- Sheldon, A. Randall, A. Charlesworth and D. Walsh, 'Popular protest and the persistence of the customary corn measures: resistance to the Winchester Bushel in the English West', in A. Randall and A. Charlesworth (eds), *Markets, Market Culture and Popular Protest in Eighteenth-century Britain and Ireland* (Liverpool 1996), pp. 25–45. The French Revolution gave a stimulus to similar advances in unifying weights and measures on the Continent but the process in France alone took over 40 years: Ken Alder, 'A revolution to measure: the political economy of the metric system in France', in M. Norton Wise (ed.), *The Values of Precision* (Princeton 1995), pp. 39–71. Witold Kula linked moves to unify measures with political revolutions more generally, stressing the connection between metrological and juridical equality. In China and Russia as well as France moves to unify measurement helped to shift the economies away from an order based on privilege to one based on law which had the added advantage of greater efficiency in administration and taxation: W. Kula, *Measures and Men* (Princeton 1986).
11. Hoppit (1996) 'Political arithmetic'; Cullen (1975) *Statistical Movement*; Porter (1995) *Trust in Numbers*, pp. viii–ix.
12. Cullen (1975) *Statistical Movement*, p. 7. The links between astronomy (both as inspiration and training) and statistical and model building ideas applied to social phenomena were very strong. Most major quantitative social theorists of the period either started out as astronomers or were involved with astronomy, most notably Adolphe Quetelet. See William J. Ashworth, 'The calculating eye: Baily, Herschel, Babbage and the business of astronomy', *British Journal of Historical Studies*, 27 (4), (1994), pp. 409–441; Ken Alder, *The Measure of All Things: The Seven-Year Odyssey and Hidden Error That Transformed the World* (New York 2002). Life tables were named after places in just the same way as astronomical tables were named after constellations.
13. Cullen (1975) *Statistical Movement*, p. 8.
14. Bayes' work was published in 1763 (two years after his death). Unfortunately his ideas proved less powerful for much of the twentieth century than those of his successors (as we shall see) even though some would argue that Bayes' theorem is a more sophisticated way of gauging the significance of experimental results. R. Matthews, 'Flukes and flaws', *Prospect*, November 1998, pp. 20–24. For the history of probability theory see Porter (1986), *Rise of Statistical Thinking* but more particularly, I. Todhunter, *A History of the Mathematical Theory of Probability* (New York 1949); Ian Hacking, *The Taming of Chance* (Cambridge 1990) and L. Daston, *Classical Probability in the Enlightenment* (Princeton 1988).
15. Daston (1988) *Classical Probability in the Enlightenment*; Hacking (1990) *The Taming of Chance*.
16. J. F. Von Bielfeld, *The Elements of Universal Erudition* (trans W. Hooper), 3 volumes (London 1770); Cullen (1975) *Statistical Movement*, p. 10.
17. Von Bielfeld (1770) *The Elements of Universal Erudition*, pp. 271–272.
18. Cullen (1975) *Statistical Movement*, p. 10.
19. Sir John Sinclair, *The Statistical Account of Scotland*, XX (Edinburgh 1798), p. xix n.
20. *Encyclopedia Britannica*, 3rd edition (Edinburgh 1797), XII, 731 quoted by Cullen (1975) *Statistical Movement*, pp. 10–11.
21. Porter (1986) *Rise of Statistical Thinking*, p. 24.
22. Cullen (1975) *Statistical Movement*, p. 11, referring to William Playfair, *The Statistical Breviary* (London 1801), p. 4.
23. Cullen (1975) *Statistical Movement*, p. 11, quoting from W. T. Brande (ed.), *A Dictionary of Science, Literature and Art* (London 1842), p. 1150.
24. Phrase used by T. M. Porter, 'Making things quantitative', *Science in Context*, 7 (3), (1994), p. 397. By this time a shift had also occurred in statistical method which became wrapped up

- in the meaning of statistics itself as the study of 'laws' of error and variation in large numbers.
- See also Mackenzie (1981) *Statistics in Britain*, p. 8 ff.
25. Cullen (1975) *Statistical Movement*, p. 14.
 26. Rickman's death and the establishment of the General Register Office in 1837 meant that categories in the 1841 census, and, gradually, in those thereafter, became more detailed and clearly defined. Interestingly, Rickman's conclusions about the timing of population change in the eighteenth century, based on the clerical data he had requested, have not been seriously undermined by later more sophisticated work. D. V. Glass, 'Some aspects of the development of demography', *Journal of the Royal Society of Arts*, 104, (1955–1956), pp. 854–869; E. A. Wrigley and R. S. Schofield, *The Population History of England and Wales, 1541–1871* (Cambridge 1981); Higgs (2004) *Information State*.
 27. This and the later version of Malthus's essay are included in E. A. Wrigley and D. Souden (eds), *The Works of Thomas Robert Malthus*, 8 volumes (London 1986). For commentary on the early development of demographic statistics see D. V. Glass, *Numbering the People* (London 1978).
 28. Census night was introduced in 1841 by Lister. E. Higgs, *Making Sense of the Census* (London 1989), p. 9.
 29. Milne developed the Carlisle tables of mortality which were thereafter often used in preference to the Northampton tables though the Select Committee on Laws respecting Friendly Societies III, 1826–1827 (p. 11) still thought fit to call for more accurate and extensive collection of facts for the construction of such tables.
 30. Porter (1995) *Trust in Numbers*, p. 36.
 31. M. Norton Wise (ed.), *The Values of Precision* (Princeton 1995), Introduction.
 32. Cullen (1975) *Statistical Movement*, pp. 135–149; Porter (1986) *Rise of Statistical Thinking*, pp. 23–39; Higgs (2004) *Information State*.
 33. Nietzsche, quoted by Porter (1994) 'Making things quantitative', p. 396.
 34. The tendency to destroy the original quantitative data once it has been processed, summarized or analysed is not unique to the nineteenth century. Materials from twentieth-century social surveys and questionnaires have often similarly been discarded. This is indicative of both carelessness and an embedded idea that data collection is done for one purpose only and cannot or should not be reanalysed later asking different questions or using other techniques. An exception to the general story of destruction of original raw data is the 1929–1931 New London Survey which has been digitized and now provides research material for a number of British economic and social historians. See for example, T. J. Hatton and R. E. Bailey, 'Poverty and the welfare state in interwar London', *Oxford Economic Papers*, 50 (4), (1998), pp. 574–606.
 35. Cullen (1975) *Statistical Movement*, p. 144 and *passim*. Tirades against Anglicanism were not characteristic of the statistical movement generally but were prominent in the influential work of G. R. Porter at the Board of Trade. Porter (1986) *Rise of Statistical Thinking*, p. 34.
 36. Cullen (1975) *Statistical Movement*, Chapter 1.
 37. Cullen (1975) *Statistical Movement*, Chapter 2.
 38. Higgs (2004) *Information State*; Cullen (1975) *Statistical Movement*, Chapter 2; E. Higgs, 'Women, occupations and work in the nineteenth century census', *History Workshop Journal*, 23 (1), (1987), pp. 59–80. For suggested corrections to the female participation rate figures see J. Humphries, 'Women and work', in J. Purvis (ed.), *Women's History, 1850–1914: An Introduction* (London 1997), pp. 85–105.

Notes

39. Crime rates, suicide rates and rates of ill health did however gradually contribute to an acknowledgement that these were features of society and not solely reflective of individual failings. Cullen (1975) *Statistical Movement*, Chapters 4 and 5. Such figures were used enthusiastically by late nineteenth-century social reformers, Fabian socialists and early collectivists and figure importantly in the origins of social work as a profession.
40. Michelle Perrot, 'Premières mesures des faits sociaux: les débuts de la statistique criminelle en France (1790–1830)', in *Pour une histoire* (1977) cited by Porter (1986) *Statistical Thinking*, p. 30.
41. Apart from Perrot, see Hacking (1990) *Taming of Chance*; Mary Poovey, *Making a Social Body: British Cultural Formation, 1830–1865* (Chicago 1985); Crook and O'Hara (2011) *Statistics and the Public Sphere*; Patrick Joyce, *The Rule of Freedom: Liberalism and the Modern City* (London 2003), pp. 24–34; Nikolas Rose, *Powers of Freedom: Reframing Political Thought* (Cambridge 1999), Chapter 6.
42. A. L. Bowley, *Wages in the United Kingdom in the Nineteenth Century* (Cambridge 1900); G. H. Wood, *The History of Wages in the Cotton Trade During the Past One Hundred Years* (1910); E. H. Phelps-Brown and S. V. Hopkins, 'Seven centuries of the price of consumables compared with builders' wage rates', in E. M. Carus Wilson (ed.), *Essays in Economic History*, 11 (London 1954), pp. 179–196; B. R. Mitchell and P. Deane, *Abstract of British Historical Statistics* (Cambridge 1962).
43. W. A. S. Hewins, revised Alon Kadish, 'James Edwin Thorold Rogers (1823–1890)', *Oxford Dictionary of National Biography* (2006), online: www.oxforddnb.com/view/article/23979, accessed 6 March 2013.
44. Rebecca Searle, 'Is there anything real about real wages? A history of the official cost of living index, 1914–1962', *Economic History Review*, 68 (1), (2015), pp. 145–166.
45. Avner Offer, "Charles Feinstein (1932–2004)" and British National Accounts, *University of Oxford Discussion Papers in Economic and Social History*, 70, (2008); Keith Tribe, 'The measurement of economic activity and the growth metric: constructing National Income in Britain 1907–41', in *The Economy of the Word* (Oxford 2015), pp. 89–107.
46. Alain Desrosieres, 'Managing the economy', in T. M. Porter and D. Ross (eds), *The Modern Social Sciences* (Cambridge 2003), pp. 553–564, p. 560.
47. S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900* (Cambridge, MA 1986); A. Desrosieres, *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, MA 1998); Porter (1986) *Statistical Thinking*; Hacking (1990) *Taming of Chance*; Crook and O'Hara (2011) *Statistics and the Public Sphere*.
48. For critical comment on the influence of Quetelet on the perception of criminal statistics in particular see Howard Taylor, 'Rationing crime: the political economy of criminal statistics since the 1850s', *Economic History Review*, 51 (3), (1998), pp. 827–851.
49. Porter (1986) *Statistical Thinking*, pp. 44–70.
50. Ian Hesketh, *The Science of History in Victorian Britain* (London 2011), especially Chapter 1 on 'The enlarging horizon: Henry Thomas Buckle's Science of History'.
51. T. M. Porter, *Karl Pearson: The Scientific Life in a Statistical Age* (Princeton, NJ 2004).
52. Stigler (1986) *History of Statistics*.
53. See S. J. Gould, *The Mismeasure of Man* (London 1981).
54. Mackenzie probably overplays this analysis a bit as his commitment to the social construction of knowledge is rather mechanistic. Mackenzie (1981) *Statistics in Britain*.

55. Pearson was committed to correlation and regression so that the degree of association between variables especially hereditability could be predicted in populations yet to be measured. For correlation and regression see Chapter 6. For discussion of this see Mackenzie (1981) *Statistics in Britain*, especially Chapter 7: ‘The politics of the contingency table’. Here Mackenzie argues that the desire for precise predictability led Pearson to extend the theory of correlation from interval variables such as height and IQ measures to nominal variables such as eye and hair colour. This is how the Pearson correlation coefficient came to be developed and accepted.
56. This belief was termed Mendelism after Gregor Mendel (1822–1884). Mendel was an Austrian monk who observed the changes produced in successive generations of pea plants by cross-fertilizing plants with different characteristics. Fisher continued these sorts of experiments and was resolutely set against attaching importance to mutationism or ‘random drift’ in evolution. Mackenzie (1981) *Statistics in Britain*, p. 191. Fisher argued in the 1930s that clergymen’s stipends should be varied according to their fertility in order to ensure reproduction of the best genes for service of the Church.
57. G. Gigerenzer, Z. Swijtink, T. M. Porter, L. Daston, J. Beatty and L. Kruger, *The Empire of Chance: How Probability Changed Science and Everyday Life* (Cambridge 1989); Daston (1988) *Classical Probability in the Enlightenment*.
58. Why did Fisher set 0.05 as the crucial dividing line for significance? Apparently for no other reason than that it was ‘convenient’. See Robert Matthews, ‘Flukes and flaws’, *Prospect*, November 1998. Acceptance of the 0.05 significance level is geared to setting a uniform standard so that results can be compared. Acceptance did not imply setting any sort of absolute standard regarding the importance and reliability of results but this is often forgotten.
59. There has been a growing popularity of the Bayesian alternative in recent years in many disciplines. Because of the complex calculations involved in Bayesian reasoning, it is only with the advent of computer-assisted calculation that the extension of the technique has been possible.
60. S. T. Ziliak and D. N. McCloskey, *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice and Lives* (Ann Arbor 2008).
61. The growth of historical relativism is usually associated with the ideas of the Italian philosopher Berndetto Croce in the 1920s. Extreme relativism is associated with post-modernism, and particularly with the writings of Hayden White. White cites Croce as a major influence. For an introduction to these developments and further reading see Anna Green and Kathleen Troup, *The Houses of History: A Critical Reader in 20th Century History and Theory* (Manchester 1999), Chapters 8–12.
62. See for example: M. Foucault, *The Order of Things* (New York 1973); R. Rorty, *Objectivity, Relativism and Truth* (Cambridge 1991); Nikolas Rose, *Governing the Soul* (London 1990).
63. A. Giddens, *The Constitution of Society* (Cambridge 1984), p. 333. I am grateful to Kevin Passmore for this quote.
64. Porter (1995) *Trust in Numbers*, p. 11; Pat Hudson, ‘Numbers and words: quantitative methods for scholars of texts’ in G. Griffin (ed.), *Research Methods for English Studies* (Edinburgh 2005), pp. 131–156. Jürgen Habermas in particular makes a distinction between the instrumentality of the empirical analytical sciences and the less problematic hermeneutic sciences: the latter are seen as oriented around meaningful communication, involving techniques such as the interpretation of documents. J. Habermas, *The Structural Transformation of the Public Sphere*, English translation (Cambridge, MA 1962). But these categories are, in practice, inextricably intertwined in what researchers do. A shared framework of meanings and assumptions, sustained by consensus and authority is present whether we are primarily quantitative in our approach to a historical research project or not.

Notes

Chapter 3 Arranging, Rearranging and Displaying Data

1. Hilary Doda, “Said Monstrous hose”: compliance, transgression and English sumptuary law to 1553, *Textile History*, 45 (2), (2014), p. 181.
2. Mass Observation was a movement founded in 1937 by Tom Harrison and Charles Madge. Its aim was to record everyday life in Britain using a panel of volunteer observers. The ‘movement’ resulted in a large and valuable collection of diaries, notebooks, surveys and interviews relating to the war period in particular. Several books were written by the mass observation team in the 1940s and 1950s, including *The Pub and the People* (London 1943). Work continued to the 1960s and was revived in the 1980s. The archive is currently housed at the University of Sussex: www.massobs.org.uk/ (accessed 20 April 2015).
3. For discussion of these problems see: www.roehampton.ac.uk/Research-Centres/Centre-for-Hearth-Tax-Research/ (accessed 16 November 2015); Margaret Spufford, *Local Historian*, 30 (4), (2000), pp. 202–221.
4. For details of the land tax as a source see M. Turner and D. Mills (eds), *Land and Property: The English Land Tax 1692–1832* (Gloucester 1986).
5. For further details of research on this question see Pat Hudson, ‘Landholding and the organisation of textile manufacture in two Yorkshire townships’ in M. Berg (ed.), *Markets and Manufactures in Early Industrial Europe* (London 1993).
6. Logarithmic and exponential scales have a different theoretical basis appropriate to different patterns of growth.
7. Settlement examinations and certificates were carried out by parishes to check whether claimants for poor relief might legitimately be the responsibility of another parish (by birth, employment or other residential qualification). K. D. M. Snell, *Annals of the Labouring Poor* (Cambridge 1985).
8. E. A. Wrigley and R. S. Schofield, *The Population History of England* (Cambridge 1981); E. A. Wrigley, R. Davies, J. Oeppen and R. S. Schofield, *English Population History from Family Reconstitution, 1580–1837* (Cambridge 1997).
9. Thomas Piketty, *Capital in the Twenty-First Century* (Harvard 2014). For more on sources for the study of inequality in Britain see Anthony B. Atkinson, *Inequality: What Can Be Done?* (Cambridge, MA 2015).
10. Frances Parkes, *Domestic Duties or Instructions to Young Married Ladies on the Management of their Households* (London 1829), digitized by Google and available at <https://archive.org/details/domesticdutieso00parkgoog> (accessed 16 November 2015).
11. Sarah Richardson, ‘Petticoat politics in eighteenth and early nineteenth-century Britain: female citizenship revealed by the digital archive’, <http://wrap.warwick.ac.uk/70990/> (accessed 10 August 2015). We are very grateful to the author for permission to use the word cloud.
12. J. Langton and R. J. Morris, *An Atlas of Industrialising Britain 1780–1914* (London 1986).

Chapter 4 Summarizing Data: Averages and Distributions

1. Thomas Piketty, *Capital in the Twenty-First Century* (Harvard 2014).
2. Reported in the *Independent*, 22 October 1999, p. 21.
3. Joseph E. Stiglitz, ‘Of the 1% by the 1% for the 1%’, *Vanity Fair*, May 2011.

4. M. Botticini, 'A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town', *Journal of Economic History*, 59 (1), (1999), pp. 104–121. For the exercise see p. 235.
5. Richard Akresh, Philip Verwimp and Tom Vundervoet, 'Civil War, crop failure and child stunting in Rwanda', *Economic Development and Cultural Change*, 59 (4), (2011), pp. 777–810.
6. Stephen L. Morgan, 'Richer and taller: stature and living standards in China, 1979–1995', *The China Journal*, 44, (2000), pp. 1–39.
7. A recent survey of research on this topic can be found in Francois Bourguignon, *The Globalization of Inequality* (Princeton 2015).
8. Paul Johnson, 'The welfare state', in R. Floud and D. N. McCloskey (eds), *The Economic History of Britain Since 1700*, volume 3, 1939–1992, 2nd edition (Cambridge 1994), pp. 284–317.
9. Recent research has however shown that studying the top 5 per cent and especially the top 1 per cent gives a rather different picture of inequality trends: Piketty (2014) *Capital*; Atkinson (2015) *Inequality*.
10. The precise way of calculating the Gini coefficient varies which results in different absolute values even when the same data are being discussed but the important thing to remember is that it is a measure used for comparison across countries or over time. As long as the same method of calculation is used, the absolute value of the Gini coefficient remains unimportant.
11. Atkinson (2015) *Inequality*.
12. M. Gutmann, S. Pullum-Piñón and T. Pullum, 'Three eras of young adult home leaving in twentieth-century America', *Journal of Social History*, 35 (3), (2002), 533–576. (See exercise on p. 237.)
13. Ideal-types are a useful aid to analysis both quantitative and non-quantitative because they enable the supposed underlying nature of real phenomena or data to be captured and discussed. Max Weber developed the use of ideal-types in analysis and this is widely adopted in social science especially in sociology.
14. Quoted by A. Desrosieres, *The Politics of Large Numbers* (Cambridge, MA 1998), p. 76.

Chapter 5 Time Series and Indices

1. D. C. Coleman, 'History, economic history and the numbers game', *Historical Journal*, 38 (3), (1995), p. 643.
2. To calculate these percentages one must place the other data values in turn over the base year original figure and multiply by 100. In other words the data values become successive numerators with the base year original figure the denominator in each case. Each resulting fraction is multiplied by 100 to convert to a percentage.
3. In this equation, the number of strikes in 1865 is the **numerator** and the number of strikes in the base year is the **denominator**. Numerator and denominator are defined in the Glossary. An excellent introduction to very simple calculations of fractions and percentages is given in R. Soloman and C. Winch, *Calculating and Computing for Social Science and Arts Students* (Buckingham 1994), Chapter 1.
4. For more discussion of this in the light of the latest historical estimates of national output see Stephen Broadberry, Bruce M. S. Campbell, Alexander Klein, Mark Overton and Bas van Leeuwen, *British Economic Growth 1270–1870* (Cambridge 2015).

Notes

5. C. Feinstein, 'Pessimism perpetuated: real wages and the standard of living in Britain during and after the industrial revolution', *Journal of Economic History*, 58 (3), (1998), p. 637.
6. Walther G. Hoffman, *British Industry 1700–1950*, trans. W. O. Henderson and W. H. Chaloner (Oxford 1955).
7. T. S. Ashton, 'The standard of life of the workers in England 1700–1830', *Journal of Economic History*, 9, supplement 1949.
8. In the exercise section after Chapter 7, later studies of living standards during the Industrial Revolution are the subject of an exercise based on articles by Peter Lindert and Jeffrey Williamson, and by Charles Feinstein. In both there is discussion of the construction of more accurate real wage figures derived from revised composite wage and cost of living indices. These indices include rents (though from a narrow evidential base) and allow for the reduced costs of cotton clothing and for more varied foodstuffs.
9. Robert C. Allen, *The British Industrial Revolution in Global Perspective* (Cambridge 2009).
10. The cycles of regular periodicity in advanced industrial economies usually last around seven years and are called juglar cycles. The longer wave cycles of innovative activity are called Kondratiev cycles after Nikolai Kondratiev, who first identified them.
11. If the sum of these deviations does not add up to 0 the 'error term' (the difference between the sum of the mean deviations and zero) should be divided equally and added to the mean deviation for each quarter.
12. For more analysis of these figures, and for discussion of the relationship between cash and credit sales and of short-term credit in the West Yorkshire textile industry more generally, see Pat Hudson, *The Genesis of Industrial Capital* (Cambridge 1986), pp. 203–207.
13. K. A. Odell and M. C. Weidenmier, 'Real shock, monetary aftershock: the 1906 San Francisco earthquake and the panic of 1907', *Journal of Economic History*, 64 (4), (2004), pp. 1002–1027, p. 1017.
14. V. Barnett, 'Soviet commodity markets during NEP', *Economic History Review*, 48 (2), (1995), pp. 329–352. See pp. 333, 350.
15. R. Price, *British Society 1680–1880* (Cambridge 1999), p. 1.

Chapter 6 Relationships Between Variables

1. Darrell Huff, *How to Lie with Statistics* (London 1973), p. 90.
2. For more on sampling and sampling bias see Chapter 7.
3. It is not always clear which way a causal connection might run especially with something like exports and imports or education and nuptiality so historians must be prepared for the possibility that the direction of causality might be different to what they expect. It is also of course possible, as we shall see, that a close relationship between variables has no causal basis whatsoever. If this is suspected there is no point in exploring it.
4. The 'politics of the contingency table' and the conflicts between Pearson and others are discussed in D. A. MacKenzie, *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge* (Edinburgh 1981), especially pp. 153–182.
5. There are many uses for the distribution. χ^2 application to contingency tables is one of the most common uses.

6. Benjamin and Kochin's work on modelling and testing the relationship between unemployment and benefits payments in inter-war Britain started a lengthy debate in the 1970s and 1980s about regional versus national-level correlations and about the degree to which the causal connections could run in the opposite direction to that posed by the authors. The debate had big exposure both inside and outside academia because it appeared to support the contemporary Thatcherite policy of 'pricing people back into jobs', that is, bringing down wages but bringing down benefits even more so that the lack of attraction of benefits would force people back to work. This of course supposed that there were jobs available and that there was no demand deficiency in the economy. See D. Benjamin and L. Kochin, 'Searching for an explanation for interwar unemployment', *Journal of Political Economy*, 87, (1979), pp. 441–478. For debate and critiques using regional measures see issues of the *Journal of Political Economy*, 1980. See also Charles Feinstein and Mark Thomas, *Making History Count: A Primer of Quantitative Methods for Historians* (Cambridge 2002), pp. 438–445, 501–503.
7. Eltjo Buringh and Jan Luiten van Zanden, 'Charting the "Rise of the West": manuscripts and printed books in Europe, a long term perspective from the sixth through the eighteenth century', *Journal of Economic History*, 69 (2), (2009), p. 427.
8. E. A. Wrigley and R. S. Schofield, *The Population History of England and Wales, 1541–1871* (Cambridge 1981).
9. J. Mokyr, 'Three centuries of population change', *Economic Development and Cultural Change*, 32, (1983), p. 190.
10. Pat Hudson, *The Genesis of Industrial Capital: A Study of the West Riding Wool Textile Industry c. 1750–1850* (Cambridge 1986), pp. 244–245.
11. Joyce Burnette, 'The wages and employment of female day labourers in English agriculture, 1740–1860', *Economic History Review*, 57 (4), (2004), pp. 664–690, pp. 668–669.
12. Federico Etro and Laura Pagani, 'The market for paintings in Italy during the seventeenth century', *Journal of Economic History*, 72 (2), (2012), pp. 423–447.
13. Tracy Dennison and Sheilagh Ogilvie, 'Does the European marriage pattern explain economic growth?', *Journal of Economic History*, 74 (3), (2014), pp. 651–693.
14. V. Barnett, 'Soviet commodity markets during NEP', *Economic History Review*, 48 (2), (1995), pp. 329–352, pp. 333, 350.
15. The countries are: Germany, Ireland, Denmark, Finland, Sweden, the UK, Australia, Switzerland, Poland, France, the Netherlands, Belgium, Czechoslovakia, Spain, Portugal, Canada, Japan, Hungary, Italy and Bulgaria.
16. The Poor Law datasets associated with this work are used extensively in examples in Feinstein and Thomas (2002) *Making History Count*, pp. 313–314, 463–480, 496–499.
17. N. F. R. Crafts, S. J. Leybourne and T. C. Mills, 'Trends and cycles in British industrial production, 1700–1913', *Journal of the Royal Statistical Society*, 152, (1989), pp. 43–60. This provoked considerable debate about the statistical elimination of regularities using filters which the more technically minded might be interested in pursuing. The most important contributions are N. F. R. Crafts and T. C. Mills, 'Modelling trends in economic history', *The Statistician*, 45 (2), (1996), pp. 153–159; David Greasly and Les Oxley, 'Endogenous growth, trend output and the industrial revolution: a reply to Crafts and Mills', *Journal of Economic History*, 57 (4), (1997), pp. 957–960.
18. In statistical parlance one needs to identify in what order the series is integrated.
19. For more on these in the context of historical research see Feinstein and Thomas (2002) *Making History Count*, pp. 312–316, 315, 321–3, 373, 439, 448.

Chapter 7 Sampling and Significance Testing

1. R. L Meek, *Figuring out Society* (London 1971), p. 72.
2. Pat Hudson, 'Correspondence and commitment: British traders' letters in the long eighteenth century', *Cultural and Social History*, 11 (4), (2014), pp. 527–553.
3. M. Anderson et al., 'The national sample from the 1851 census of Great Britain', *Urban History Newsletter* (1977), pp. 55–59.
4. Their detailed workings are explained at www.nappdata.org/napp/resources/publications_pdf/1881gbsample.pdf (accessed 14 September 2015).
5. Huw Beynon, *Working for Ford* (London 1973), p. 14.
6. L. Weatherill, *Consumer Behaviour and Material Culture in Britain, 1660–1760* (London 1988).
7. Mark Overton, Jane Whittle, Darren Dean and Andrew Hahn, *Production and Consumption in English Households, 1600–1750* (Abingdon 2004).
8. For discussion of how the authors do this see p. 218.
9. Paul Thompson, *The Edwardians* (St Albans 1977).
10. Emma Griffin, *Liberty's Dawn: A People's History of the Industrial Revolution* (Yale 2013).
11. For example Julian Hoppit, *Risk and Failure in English Business 1700–1800* (Cambridge 1987).
12. C. McCloed and A. Nuvolari, 'The pitfalls of prosopography: inventors in the "Dictionary of National Biography"', *Technology and Culture*, 47 (4), (2006), pp. 747–776.
13. See www.northwestfamilybusiness.arts.manchester.ac.uk/Intro.aspx (accessed 10 November 2015).
14. R. Floud, *An Introduction to Quantitative Methods for Historians* (London 1973; 2nd edition 1979), Chapter 8. His example derived from parish register sources but these are difficult to use for this purpose as so few give marriage ages. These generally have to be calculated by comparing the date of baptism (if it can be found, especially for the bride) with the date of marriage, allowing for the addition of an average birth/baptism interval. Civil Registration data gives marriage ages directly and, generally, much more accurately.
15. Census enumerators' transcripts for 1881 and for many other nineteenth-century censuses are now available in machine-readable form. Details can be found at <http://hds.essex.ac.uk/>. See Matthew Woppard, 'Creating a machine-readable version of the 1881 Census of England and Wales', in C. Harvey and J. Press (eds), *Databases in Historical Research* (Basingstoke 1996), pp. 98–101.
16. The Guinness Company prohibited employees from publishing their work, so Gosset published under the pseudonym 'Student'. The *t*-distribution test is thus sometimes called the Student's *t*-test. For debates concerning Gosset's test and its interpretation by R. A. Fisher (mistaken in their view), see Stephen T. Ziliak and Deirdre McCloskey on 'The cult of statistical significance' at www.press.umich.edu/pdf/9780472070077-fm.pdf (accessed 30 September 2015).
17. Throughout these discussions of sampling and significance testing it is well to bear in mind the difference between statistical significance and historical significance and the debates covered in Stephen T. Ziliak and Deirdre McCloskey: www.deirdremccloskey.com/docs/jsm.pdf; and in 'The cult of statistical significance' at www.press.umich.edu/pdf/9780472070077-fm.pdf (both accessed 30 September 2015).

18. G. R. Boyer and T. J. Hatton, 'Migration and labour market integration in late nineteenth century England and Wales', *Economic History Review*, 50 (4), (1997), pp. 697–734.
19. M. Botticini, 'A loveless economy? Intergenerational altruism and the marriage market in a Tuscan town, 1415–1436', *Journal of Economic History*, 59 (1), (1999), pp. 104–121.
20. B. Gupta, 'Where have all the brides gone? Son preference and marriage in India over the twentieth century', *Economic History Review*, 67 (1), (2014), pp. 1–24.
21. YEAR dummy is equal to 1 for marriages after 1427; AGE DIFFERENCE is the groom's age minus the bride's age; the omitted dummy is AGRAGRDOWN; AGRUPDOWN is equal to 1 when a bride from a peasant household (AGR) marries a non-peasant groom (UP) with lower wealth (DOWN). Analogously, AGRUPUP equals 1 when a bride from a peasant household (AGR) marries a non-peasant groom (UP) with higher wealth (UP); NONAGRDOWNDOWN (NONAGRDOWNUP) are similarly defined for non-peasant brides marrying peasant grooms; NONAGRNONAGRDOWN (NONAGRNONAGRUP) are similarly defined for non-peasant brides marrying non-peasant grooms; RESIDENCE dummy is equal to 1 if the bride's and/or the groom's households lived in the town of Cortona, and is equal to 0 if both households lived in the countryside.

Chapter 8 Economic History and Econometric History

1. See Chapter 2.
2. For recent analyses of the rise of economic history worldwide, and its varying nature, see Francesco Boldizzoni and Pat Hudson (eds), *The Routledge Handbook of Global Economic History* (London 2016).
3. D. N. McCloskey, *Econometric History* (London 1987), p. 44.
4. Herbert Spencer quoted by Alfred H. Conrad and John R. Meyer, *Studies in Econometric History* (London 1965), p. 3.
5. The *Journal of Economic History* and *Economic History Review* are the major journals in the United States and the UK, respectively, that cover economic history more generally, so it is instructive to note the changing proportion of articles that use statistical method as an index of first the growth and then the subsequent established importance of econometric techniques during recent decades, as in Robert Whaples, 'The quantitative history of the *Journal of Economic History* and the cliometric revolution', *Journal of Economic History*, 51 (2), (1991), pp. 289–301.
6. This American journal ran from 1949 to 1958 as *Explorations in Entrepreneurial History*, first series volumes 1–10 and second series (1963–1969) volumes 1–6. Since 1969 it has been called *Explorations in Economic History*, starting with volume 7. Entrepreneurial history experienced considerable early application of econometric techniques and the journal now specializes in the publication of econometric history.
7. For a recent review of the impact in the USA see Naomi Lamoreaux, 'Beyond the old and the new: economic history in the USA', in Francesco Boldizzoni and Pat Hudson (eds), *The Routledge Handbook of Global Economic History* (London 2016), pp. 35–54.
8. By formal economic models one here means the construction of a set of interlocking economic relationships that can be hypothesized to exist on the basis of the first principles of economic theory, generally neoclassical economic theory. From such models, often constructed using dedicated computer software, the theoretical impact of change in any one variable can be calculated.

Notes

9. This has narrowed the scope of academic economics as a social science whilst broadening it as a field of mathematical or statistical enquiry. For an early critique of this development see Benjamin Ward, *What's Wrong with Economics?* (London 1972). For an amusing account of the intellectual college of economics and of the mindset and habits of economists as a tribe (that still seems relevant today) see Axel Leijonhufud, 'Life among the econ', *Western Economic Journal*, (1973), pp. 327–337. On the narrowing of economics as a discipline see B. Caldwell, *Beyond Positivism: Economic Methodology in the Twentieth Century* (Boston, MA 1982); Geoffrey Hodgson, *How Economics Forgot History: The Problem of Historical Specificity in Social Science* (London 2001).
10. R. W. Fogel and G. R. Elton, *Which Road to the Past? Two Views of History* (London 1983), p. 2.
11. The account given and the historians interviewed in John S. Lyons, Louis P. Cain and Samuel H. Williamson (eds), *Reflections on the Cliometrics Revolution. Conversations with Economic Historians* (London 2008) suggest the nature of the overlap between cliometric and econometric approaches.
12. E. A. Wrigley and R. S. Schofield, *The Population of England and Wales, 1541–1871* (London 1981).
13. For a fairly technical survey and assessment of social savings methods see T. Leunig, 'Social savings', *Journal of Economic Surveys*, 24 (5), (2010), pp. 775–800.
14. Niall Ferguson (ed.), *Virtual History: Alternatives and Counterfactuals* (London 2011); Andrew Roberts, *What Might Have Been. Imaginary History from Twelve Leading Historians* (London 2005). For a less sympathetic treatment of the approach see Richard J. Evans, *Altered Pasts: Counterfactuals in History* (New York 2014).
15. This history is admirably traced and documented by Mary Morgan, *The History of Econometric Ideas* (Cambridge 1990). For a summary accounts see McCloskey (1987), *Econometric History*. For the USA story see Naomi Lamoreaux, 'Economic history and the cliometric revolution', in A. Molho and G. S. Wood (eds), *Imagined Histories: American Historians Interpret the Past* (Princeton 1998), pp. 59–84.
16. C. H. Lee, *The Quantitative Approach to Economic History* (London 1977), p. 5.
17. Morgan (1990), *History of Econometric Ideas*, p. 6.
18. Morgan (1990), *History of Econometric Ideas*, Introduction.
19. These are all briefly described in McCloskey (1987), *Econometric History*.
20. The flavour of this confidence can be gained from R. Fogel and S. Engerman (eds), *The Reinterpretation of American Economic History* (New York 1971).
21. P. Temin (ed.), *New Economic History* (Harmondsworth 1973), p. 8.
22. D. N. McCloskey, 'The achievements of the cliometrics school', *Journal of Economic History*, 38, (1978), pp. 13–28.
23. R. Solow, 'Economics: is something missing?', in W. N. Parker (ed.), *Economic History and the Modern Economist* (Oxford 1986), p. 26, quoted by N. F. R. Crafts, 'Economic history', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: A Dictionary of Economics*, volume 2 (London 1987), p. 38.
24. A. Fishlow, *American Railroads and the Transformation of the Antebellum Economy* (Cambridge, MA 1965).
25. G. R. Hawke, *Railways and Economic Growth in England and Wales* (Oxford 1970); J. Metzer, 'Railroads in Tsarist Russia: direct gains and implications', *Explorations in Economic History*, 13 (1), (1976), pp. 85–111.

26. Such problems are summarized well by P. K. O'Brien, *The New Economic History of the Railways* (London 1977) and in Lee (1977) *The Quantitative Approach to Economic History*, Chapter 4. See also C. M. White, 'The concept of social saving in theory and practice', *Economic History Review*, 29 (1), (1976), pp. 82–100. Rainer Fremdling's study of railways in Germany is perhaps the best of the many emulators of Fogel's method: *Eisenbahnen und deutsches Wirtschaftswachstum, 1840–1870* (Dortmund 1985).
27. Some of these issues are discussed in Lee (1977) *The Quantitative Approach to Economic History*. Outside of the railways the only major counterfactual study of innovation in Britain at the time was von Tunzelman's research on the steam engine which came to similar conclusions about the continued productiveness of the older technologies; G. N. von Tunzelman, *Steam Power and British Industrialization to 1860* (Oxford 1978). Similar but less complex innovation models were applied to other innovations; for example, coke smelting the mechanical reaper and the coal cutting machine. In each case the neoclassical model proved capable of sharpening the questions for discussion of highlighting the importance of comparative costs in innovation decisions: D. Greasley, 'The diffusion of machine cutting in the British coal industry, 1902–1938', *Explorations in Economic History*, 19 (3), (1982), pp. 246–268; Paul A. David, *Technical Choice, Innovation and Economic Growth* (Cambridge 1975); C. K. Hyde, 'Technological change in the British wrought iron industry 1756–1815: a reinterpretation', *Economic History Review*, 27 (2), (1974), pp. 190–206. More recently Nicholas Crafts has reevaluated the contribution of steam power as a general purpose technology using updated growth accounting tools. He finds that it contributed little before 1830 to growth in Britain and only with high pressure steam after 1850 did the new technology reach its full potential. 'Steam as a general purpose technology: a growth accounting perspective', *Economic Journal*, 114 (495), (2004), pp. 338–351.
28. The second volume was entirely devoted to the statistical appendices giving critics the research material that they needed to establish their careers.
29. Paul A. David, H. Gutman, R. Sutch, P. Temin and G. Wright, *Reckoning with Slavery: A Critical Study of the Quantitative History of American Negro Slavery* (New York 1976), p. 341.
30. D. N. McCloskey, 'Did Victorian Britain fail?', *Economic History Review*, 23 (3), (1970), pp. 446–459.
31. The key research here was D. N. McCloskey, *Economic Maturity and Entrepreneurial Decline: British Iron and Steel, 1870–1913* (Cambridge, MA 1973) and L. G. Sandberg, *Lancashire in Decline* (Columbus, OH 1974).
32. For example, Kennedy demonstrated that market failures reduced the check on inefficient management and Crafts and Thomas argued that comparative advantage may have been confined to low wage labour intensive export industries whilst the investment in human capital was also suboptimal: N. F. R. Crafts and M. F. Thomas, 'Comparative advantage in UK manufacturing trade, 1910–1935', *Economic Journal*, 96, (1986), pp. 629–645; W. P. Kennedy, 'Foreign investment, trade and growth in the United Kingdom, 1870–1913', *Explorations in Economic History*, 11, (1974), pp. 415–444.
33. For summary coverage of this debate and all the relevant references see Hudson, *The Industrial Revolution* (London 1992), Chapter 2.
34. J. G. Williamson, *Coping with City Growth during the Industrial Revolution* (Cambridge 1990).
35. G. R. Boyer, *An Economic History of the English Poor Law 1750–1850* (Cambridge 1990).
36. J. G. Williamson, *Late Nineteenth Century American Development: A General Equilibrium History* (Cambridge 1974).
37. Lee (1997) *The Quantitative Approach to Economic History*, pp. 91–92.

Notes

38. In this sense economics has a natural affinity with statistics. This is no accident because, like biometrics and the other disciplines of social science, it evolved alongside the development of statistical theory geared to ideas about the regularities of large numbers. See A. Desroisieres, *The Politics of Large Numbers: A History of Statistical Reasoning* (Cambridge, MA 2002).
39. M. Douglas and B. Isherwood, *The World of Goods: Towards an Anthropology of Consumption* (Harmondsworth 1980), Chapter 1 'Why people want goods' remains the best simple thing to read on the shortcomings of utility theory.
40. D. N. McCloskey, 'The economics of choice', in T. G. Rawski (ed.), *Economics and the Historian* (London 1996), p. 143. McCloskey also mentions here that it was reported that 10 per cent of French companies currently make use of astrologers.
41. It is interesting that demographic transition from high to low fertility that occurred in many 'advanced' countries between 1870 and the 1930s has been seen as evidence that it is at least arguable that the neoclassical optimization model is appropriate in demographic analysis for the past 100 years of West European fertility (if not earlier) when group rules appear to have predominated: E. A. Wrigley, 'Fertility strategy for the individual and the group', in C. Tilly (ed.), *Historical Studies in Changing Fertility* (Princeton 1978).
42. D. C. Coleman, 'History, economic history and the numbers game', *Historical Journal*, 38 (3), (1995), p. 643.
43. Marshall Sahlins, *Stone Age Economics* (London 1974), p. 14.
44. For more on the difficulties that economists have in theorizing gift-giving and reciprocity see Avner Offer, 'Between the gift and the market: the economy of regard', *Economic History Review*, 50 (3), (1997), pp. 450–476.
45. Coleman (1995) 'History, economic history and the numbers game', p. 643.
46. Work by Feinstein and by Crafts, Harley and their critics (described above) have been central here. C. Feinstein, *National Income, Expenditure and Output of the United Kingdom, 1855–1965* (Cambridge 1972); also E. A. Wrigley and R. S. Schofield, *The Population of England and Wales, 1541–1871* (London 1981).
47. Some of the earlier developments were explored in a highly accessible manner in John Maloney (ed.), *What's New in Economics* (Manchester 1992). For more recent summaries see D. Greasley and L. Oxley (eds), Special Issue of *Journal of Economic Surveys*, 24 (5), (2010); Nicholas Crafts, 'Solow and growth accounting: a perspective from quantitative economic history', *History of Political Economy*, 41 (1), (2009), pp. 200–220.
48. For a flavour of the contributions from sociologists and geographers that are not always compatible with quantitative or neoclassical theorizing, see R. Swedberg (ed.), *Economic Sociology* (1996), Part 2 and R. Lee and J. Wills (eds), *Geographies of Economies* (London 1997).
49. N. F. R. Crafts, 'Exogenous or endogenous growth: the industrial revolution reconsidered', *Journal of Economic History*, 55 (4), (1995), pp. 745–772.
50. The key text on this is B. Hillier, *The Economics of Asymmetric Information* (Basingstoke 1997).
51. M. J. Olney, 'When your word is not enough: race, collateral and household credit', *Journal of Economic History*, 58 (2), (1998), pp. 408–431. This article is used as an exercise on p. 239.
52. Carlos concluded that the Royal Africa company contracts, which included a bond and high pay, were well designed to mitigate moral hazard and that the failure of the company cannot be attributed to inefficiencies in this respect: A. M. Carlos, 'Bonding and the agency problem: evidence from the Royal Africa Company, 1672–1691', *Explorations in Economic History*, 31, (1994), pp. 313–315. The Hudson's Bay Company understood the agency problem and had strategies in place to minimise opportunistic behaviour: A. M. Carlos and S. Nicholas, 'Agency

- problems in early chartered companies: the case of the Hudson's Bay Company', *Journal of Economic History*, 50 (4), (1990), pp. 853–875.
53. For a further example of historical application of principal-agency theory see D. Sunderland, 'Principals and agents: the activities of the Crown agents for the colonies, 1880–1914', *Economic History Review*, 52 (2), (1999), pp. 284–306. For an introduction to the theory see Maloney (1992), *What's New*, pp. 61–65, 116–117.
54. The path-breaking historical work here was D. C. North and R. P. Thomas, *The Rise of the Western World: A New Economic History* (Cambridge 1973). For North's later position see *Understanding the Process of Economic Change* (Princeton 2005). For the place of institutions in economic theory see O. E. Williamson, *The Economic Institutions of Capitalism* (New York 1985). In relation to economic history: Avner Greif, 'Micro theory and recent developments in the study of economic institutions through economic history', in D. M. Kreps and K. F. Wallis (eds), *Advances in Economics and Econometrics: Theory and Applications*, volume 2 (Cambridge 1997) pp. 79–113.
55. G. Grabher (ed.), *The Embedded Firm: The Socio-economics of Industrial Networks* (London 1993).
56. J. Humphries, 'Towards a family friendly economics', *New Political Economy*, 3 (2), (1998), p. 237. There is now a large literature of feminist critiques of neoclassical economics. See for example, J. Humphries (ed.), *Gender and Economics* (Aldershot 1995); Nancy Folbre, *Valuing Children: Rethinking the Economics of the Family* (Cambridge, MA 2010); Francesca Bettio and Alina Verashchagina, *Frontiers in the Economics of Gender* (London 2008). For Becker's approach see G. S. Becker, *The Economic Approach to Human Behaviour* (Chicago 1976).
57. R. Nozick, *The Nature of Rationality* (Princeton 1993).
58. As an economist would express it: discounting in economic decision-making is not exponential but hyperbolic. The path-breaking work in developing the theory of myopic choice is found in G. Ainslie, *Pico-economics* (Cambridge 1992). See Maloney (1992) *What's New in Economics*, pp. 264–266, 302–304.
59. K. J. Arrow, 'Economic theory and the hypothesis of rationality', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: A Dictionary of Economics*, volume 2 (London 1987), pp. 69–74.
60. For accessible introductory discussion of time inconsistency, rationality and the force of rational expectations see Maloney (1992) *What's New in Economics*.
61. David Greasley and Les Oxley, 'Clio and the economist: making historians count', *Journal of Economic Surveys*, 24 (5), (2010), pp. 755–774.
62. Perhaps most obviously felt in modelling institutions, endogenous growth theory, directed technological change and economic geography.
63. Just to give one example, the use of Bayesian alternatives to the Pearson probability paradigm has only recently become possible in many applications because of the degree to which computer software can assist in the discounting of chance. Bayesian approaches appear to be taking off across a range of academic subjects.
64. The work successively of Deane and Cole, Feinstein, Crafts and Harley were the most prominent: P. Deane and W. A. Cole, *British Economic Growth 1688–1959* (Cambridge 1966); Feinstein (1972) *National Income, Output and Expenditure of the UK 1855–1965*; N. F. R. Crafts, *British Economic Growth during the Industrial Revolution* (Oxford 1985); N. F. R. Crafts and C. K. Harley, 'Output growth and the British industrial revolution: a restatement of the Crafts–Harley view', *Economic History Review*, 45, (1992), pp. 703–730.
65. The new figures are principally associated with the research of Crafts and can be found in his *British Economic Growth* though some of the figures have been adjusted in subsequent

Notes

- journal articles. For a critique of such estimations see J. Hoppit, 'Counting the industrial revolution', *Economic History Review*, 43, (1990), and Hudson (1992) *The Industrial Revolution*, Chapter 2.
66. See for example Crafts (1985) *British Economic Growth*; R. C. O. Matthews, C. H. Feinstein and J. C. Odling Smee, *British Economic Growth 1856–1973* (Oxford 1982).
67. For example S. Rebelo, 'Long run policy analysis and long run growth', *Journal of Political Economy*, 99, (1991), pp. 500–521; P. Romer, 'Increasing returns and long run growth', *Journal of Political Economy*, 94, (1986), pp. 1002–1037. For discussion see Nicholas Crafts, 'Solow and growth accounting: a perspective from quantitative economic history', *History of Political Economy*, 41 (1), (2009), pp. 200–220; Crafts (1995) 'Exogenous or endogenous growth? The industrial revolution reconsidered', pp. 745–772.
68. For example Kevin O'Rourke and J. G. Williamson, *Globalization and History: The Evolution of the Nineteenth Century Atlantic Economy* (Cambridge, MA 1999); J. Baten and J. L. Van Zanden, 'Book production and the onset of modern economic growth', *Journal of Economic Growth*, 13, (2008), pp. 217–235.
69. N. F. R. Crafts and A. J. Venables, 'Globalization in history: a geographical perspective', in Michael D. Bordo, Alan M. Taylor and J. G. Williamson (eds), *Globalization in Historical Perspective* (Chicago 2003); Stanley Engerman and Kenneth Sokoloff, *Economic Development in the Americas: Endowments and Institutions* (Cambridge 2012).
70. Daron Acemoglu and James A. Robinson, *Why Nations Fail. The Origins of Power, Prosperity and Poverty* (London 2012).
71. Kenneth Pomeranz, *The Great Divergence. China, Europe and the Making of the Modern World Economy* (Princeton 2000).
72. The Great Divergence in wages and prices was the subject, for example, of a special issue of the *Economic History Review*, 64 (S1), (2011). An article from this supplement, R. C. Allen, J. P. Bassino, D. Ma, C. Moll-Murata and J. L. Van Zanden, 'Wages, prices and living standards in China 1738–1925 compared with Europe, India and Japan' is included as an exercise on p. 233.
73. See www.rug.nl/research/ggdc/. The centre grew out of the work of Angus Maddison who was the first economist to try to map comparative GDP per capita figures for large parts of the globe and for long time periods.
74. The key example again here is Robert Allen, *The Industrial Revolution in Global Perspective* (Cambridge 2009). See pp. 140–141 in this volume. Daron Acemoglu, 'Directed technical change', *Review of Economic Studies*, 69, (2002), pp. 781–809; Daron Acemoglu, 'When does Labor scarcity encourage innovation?', *Journal of Political Economy*, 118 (6), (2010), pp. 1037–1078.
75. Diane Coyle, *GDP: A Brief but Affectionate History* (Princeton 2014), pp. 106ff; Morten Jerven, *Poor Numbers: How We are Misled by African Development Statistics and What to Do about It* (Ithaca 2013); Dirk Philipsen, *The Little Big Number: How GDP Came To Rule The World and What To Do About It* (Princeton 2015).
76. GDP in many economies is also less appropriate for growth measurement than GNP which is net of the earnings of foreign direct investment.
77. The difficulties of applying a transnationally acceptable standard for GDP are discussed by Coyle (2014) *GDP: A Brief History*.
78. See for example Alejandra Irigoin, 'The end of the silver era: the consequences of the breakdown of the Spanish peso standard in China and the United States, 1780s–1850s', *Journal of World History*, 20 (2), (2009), pp. 207–243.
79. Nicholas Crafts, 'Economic history', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave: A Dictionary of Economics*, volume 2 (London 1987), p. 39.

80. This has been regarded as important in the hypothetico-deductive methodology of the sciences and social sciences and was stressed as such by Karl Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge* (London 1963).
81. Crafts (1987) 'Economic history', p. 39.
82. Quoted in R. Floud (1987), 'Cliometrics', in J. Eatwell, M. Milgate and P. Newman (eds), *The New Palgrave*, p. 253.
83. Coleman (1995) 'History, economic history and the numbers game', p. 641.
84. As exemplified by the Clio-infra project at Groningen (www.clio-infra.eu/, accessed 3 January 2016) or by Piketty (2014) *Capital in the Twenty-first Century*.

Chapter 9 Historical Research, Computing and the Digital Revolution

1. As we shall discuss later in the chapter, there are other dangers with using so-called 'Big Data' arising partly from the fact that it has been collected and stored with varying levels of accuracy, professionalism and concern for compatibility and comparability across datasets. See for example the critique posed by Tim Hitchcock in 'Confronting the digital or how academic history writing lost the plot', *Cultural and Social History*, 10 (1), (2013), pp. 9–23.
2. Practical guidance in the use of computer software such as spreadsheets is readily available in other sources and not provided here. There are straightforward online guides and manuals provided with different software packages. In addition see M. J. Lewis and R. Lloyd-Jones, *Using Computers in History: A Practical Guide to Data Presentation, Analysis and the Internet* (London 1996) which is dated but still surprisingly effective; Sonja Cameron and Sarah Richardson, *Using Computers in History* (Basingstoke 2005); and, for more advanced statistical processing, Charles Feinstein and Mark Thomas, *Making History Count: A Primer in Quantitative Methods for Historians* (Cambridge 2009).
3. In addition to the six broad types outlined here there are of course also many other sorts of software that might assist the historian if working with non-quantitative evidence such as video sources, photography, artefacts requiring technical or scientific analysis and so on, but we concentrate here only on computer applications that involve some form of quantification.
4. As evidenced by the content of the journal *History Microcomputer Review* (1985–1990), later *History Computer Review* (1990–2002).
5. For example, .xls (the binary file format of Excel) or .csv ('comma separated values') format that can be opened into a spreadsheet program.
6. D. Gilbert and H. Southall, 'A database of nineteenth century British labour markets', *Journal of Historical Geography*, 16 (2), (1990), pp. 276–293; E. Rind, A. Jones and H. Southall, 'How is post industrial decline associated with the geography of physical activity?', *Social Science and Medicine*, 104, (2014), pp. 88–97.
7. For more on this see articles in S. W. Baskerville, Pat Hudson and R. J. Morris (eds), *History and Computing Special Issue: Record Linkage*, 4, (1992), especially S. A. King, 'Record linkage in a proto-industrial community'; Pat Hudson, 'A new history from below: computers and the maturing of local and regional history', *The Local Historian*, 25 (4), (1995). On the original process of computer-aided family reconstitution, as carried out by the Cambridge Group see E. A. Wrigley and R. S. Schofield, 'Nominal record linkage by computer and the logic of family reconstitution', in E. A. Wrigley (ed.), *Identifying People in the Past* (London 1973).
8. For the nature of the family reconstitution software and research accomplished using it see E. A. Wrigley, R. S. Davies, J. E. Oeppen and R. S. Schofield, *English Population History from*

Notes

- Family Reconstitution, 1580–1837* (Cambridge 1997); for more on probate data, storage and analysis see Darren Dean, Andrew Hann, Mark Overton and Jane Whittle, *Production and Consumption in English Households, 1600–1750* (London 2012).
9. For earlier discussion of nominal record linkage see Baskerville, Hudson and Morris (1992) *History and Computing*, 4. On the shortcomings of expert systems and the alternative research into neuron networks see Roger Penrose, *The Emperor's New Mind: Concerning Computers, Minds and the Laws Of Physics* (Oxford 1990).
 10. There were two excellent early guides to database design and management for historians : C. Harvey and J. Press, *Databases in Historical Research* (Basingstoke 1996) and L. Weatherill and V. Hemingway, *Using and Designing Databases for Academic Work. A Practical Guide* (Newcastle 1994). The problems of entry and coding and the need to work towards a standard to facilitate the comparability of databases is discussed in D. I. Greenstein, *Modelling Historical Data: Towards A Standard For Encoding And Exchanging Machine Readable Texts* (St Katharinen 1991). The pitfalls of lack of compatibility within many large transnational databases and with datasets built up over a period of time in which technology has changed, is discussed in Hitchcock, 'Confronting the digital'.
 11. Early applications of this sort, and their implications for history, are discussed by James E. Everett, in 'Annual review of information technology developments for economic and social historians, 1997', *Economic History Review*, 51 (2), (1998).
 12. These latter are called Boolean operators after George Boole, who first distinguished them. See Ian Dey, *Qualitative Data Analysis* (London 1993), p. 174.
 13. The main work on concordance has, for obvious reasons, been done by literary scholars as with the Digital Miscellanies Project at the University of Oxford that has contributed to our understanding of poetic culture in the eighteenth century: <http://digitalmiscellaniesindex.org/about/> (accessed 5 August 2015).
 14. The state of the art of this area of historical computing in 2008 was outlined and debated in a special double issue of the *International Journal of Humanities and Arts Computing*, 2 (1–2), (2008).
 15. For an early introduction to the principles of GIS, many of which still apply, see D. J. Maguire, M. F. Goodchild and D. W. Rind (eds), *Geographical Information Systems: Principles and Applications*, 2 volumes (New York 1991); more recent developments are partly captured in Ian N. Gregory and Alistair Geddes (eds), *Toward Spatial Humanities. Historical GIS and Spatial History* (Bloomington 2014).
 16. H. Southall, 'Applying historical GIS beyond the academy: four use cases of the Great Britain HGIS'; and Robert M. Schwartz and Thomas Thevenin, 'Railways and agriculture in France and Great Britain, 1850–1914', both in Ian N. Gregory and Alistair Geddes (eds), *Toward Spatial Humanities: Historical GIS and Spatial History* (Bloomington 2014), pp. 92–117, 4–34.
 17. For surveys of the impact of GIS in history, and the ways it can be used, see: A. K. Knowles (ed.), 'Historical GIS: the spatial turn in social science history', Special Issue of *Social Science History*, 24 (3), (2000); D. J. Bodenhamer, J. Corrigan and T. M. Harris, *Spatial Humanities, GIS and the Future of Humanities Scholarship* (Bloomington 2010); Gregory and Geddes (2014) *Toward Spatial Humanities*.
 18. I. N. Gregory, *A Place in History: A Guide to Using GIS in Historical Research* (Oxford 2003); F. Harvey, *A Primer of GIS: Fundamental Geographic and Cartographic Concepts* (New York 2008); I. Heywood, S. Cornelius and S. Carver, *An Introduction to Geographical Information Systems* (4th edition, Harlow 2012).
 19. 'A Vision of Britain through Time' is linked to the main Great Britain Historical GIS, the establishment of which is carefully described by Humphrey Southall in a series of three

- articles: 'Rebuilding the Great Britain Historical GIS, Part 1: building an indefinitely scalable statistical database', *Historical Methods*, 44, (2011), pp. 149–159; 'Rebuilding the Great Britain Historical GIS, Part 2: a geospatial ontology of administrative units', *Historical Methods*, 45, (2012), pp. 119–134; 'Rebuilding the Great Britain Historical GIS, Part 3: integrating qualitative content for a sense of place', *Historical Methods*, 47, (2014), pp. 31–44. See also H. Southall (2014) 'Applying historical GIS beyond the academy'.
20. These sorts of projects are explored in recent issues of the *International Journal of Humanities and Arts Computing* and in Gregory and Geddes (2014) *Toward Spatial Humanities*.
 21. H. M. Dunsford and S. J. Harris, 'Colonization of the wasteland of County Durham, 1100–1400', *Economic History Review*, 56 (1), (2003), pp. 34–56.
 22. B. M. S. Campbell and K. Bartley, *England on the Eve of the Black Death: An Atlas of Lay Lordship, Land and Wealth, 1300–49* (Manchester 2006).
 23. Michael Cook was the leading figure in this development. See for example, Margaret Proctor and Michael Cook, *Manual of Archival Description* (3rd edition, Farnham 2000).
 24. See www.data-archive.ac.uk/ (accessed 3 January 2016).
 25. ICPSR is at www.icpsr.umich.edu/.
 26. An up-to-date directory of a large selection of electronically available data and resources for historians can be found in the appendix to this volume at: www.bloomsbury.com/history-by-numbers.
 27. Such as the Masters and PhD programmes in Big Data Analytics at the University of Pittsburgh, www.ischool.pitt.edu/ist/degrees/specializations/big-data.php (accessed 6 February 2016).
 28. Patrick Manning, *Big Data in History: A World Historical Archive* (London 2013). See also Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (London 2015).
 29. www.ggdcc.net/maddison/ (accessed 6 February 2016).
 30. www.pop.umn.edu/ and www.ipums.org/ (accessed 6 February 2016).
 31. www.campop.geog.cam.ac.uk/ and www.censusmosaic.org/ (accessed 6 February 2016).
 32. See www.nappdata.org/ (accessed 6 February 2016).
 33. See: Stephen T. Ziliak and Deirdre McCloskey, *The Cult of Statistical Significance: How The Standard Error Costs Us Jobs, Justice and Lives* (Ann Arbor 2008).
 34. Matthew G. Kirschenbaum, *Track Changes: A Literary History of Word Processing* (Cambridge, MA 2016) is likely to prove seminal in identifying the impact of word processing upon writing.
 35. For an introduction to debates about the importance of language and discourse in determining the parameters of knowledge and communication see Keith Jenkins, *The Postmodern History Reader* (London 1997); A. Munslow, *Deconstructing History* (London 1997). On Bakhtin, whose ideas are important for a range of historical work see P. Morris (ed.), *The Bakhtin Reader* (London 1994).
 36. See www.northwestfamilybusiness.arts.manchester.ac.uk/ (accessed 11 February 2016).
 37. For further discussion of the potentialities in this respect and a survey of the ways in which computer-aided research was contributing to change in regional and local history in the early 1990s, see Hudson (1995) 'A new history from below'.
 38. See <http://practitioners.exeter.ac.uk/> and <https://www.ucl.ac.uk/lbs/> (accessed 11 February 2016).
 39. See www.londonlives.org (accessed 15 September 2015).

Notes

40. Yuzuru Isoda, Akihiro Tsukamoto, Yoshihiro Kosaka, Takuya Okumura, Masakazu Sawai, Keijo Yano, Susumu Nakata and Satoshi Tanaka, 'Reconstruction of Kyoto of the Edo period based on arts and historical documents: 3D urban model based on historical GIS data', *International Journal of Humanities and Arts Computing*, 3 (1–2), (2009), pp. 21–38; Julian Hallam and Les Roberts, 'Mapping the city in film', and Niall Cunningham, 'Troubled geographies: a historical GIS of religion, society and conflict in Ireland since the Great Famine', both in Gregory and Geddes (2014) *Toward Spatial Humanities*, pp. 62–88, 143–171.
41. Jane Humphries, *Childhood and Child Labour in the British Industrial Revolution* (Cambridge 2010).
42. See R. H. Steckel, 'Heights and human welfare: recent developments and new directions', *Explorations in Economic History*, 46 (1), (2009), pp. 1–23.
43. R. H. Steckel and Jerome C. Rose (eds), *The Backbone of History: Health and Nutrition in the Western Hemisphere* (New York 2002).
44. Phil Withington, 'The semantics of "peace" in early modern England', *Transactions of the Royal Historical Society*, 23, (2013), pp. 127–153.
45. For a useful account of procedures in qualitative data analysis, though not related to historical research, see Dey (1993) *Qualitative Data Analysis*.
46. Google Books alone has digitized several million of an estimated 1.3 billion volumes.
47. Hitchcock (2013) 'Confronting the digital', p. 18.
48. Hitchcock discusses this and other biases and problems occasioned by the digital revolution in 'Confronting the digital'.
49. Hitchcock (2013) 'Confronting the digital', p. 9.
50. This term was coined by Clifford Geertz to denote a form of enquiry dependent upon minute observation of traces and manifestations of local culture which reflect wider aspects of society. This was a counterpoise to a positivism dependent upon exhaustive 'scientific' enquiry, often at the level of nations or larger units, and using scientific methodologies of cause and effect such as hypothesis testing. Geertz's discussion of the methodological issues involved in the growing emphasis upon linking action to its sense rather than phenomena to their cause is brilliantly described in *The Interpretation of Cultures* (New York 1973) and in *Local Knowledge: Further Essays in Interpretive Anthropology* (New York 1983).
51. E. Le Roy Ladurie, *The Peasants of Languedoc* (Champaign 1976), p. 4.
52. There are many pitfalls in using the land tax. These are explored in M. Turner and D. Mills (eds), *Land and Property: The English Land Tax, 1692–1832* (Gloucester 1986).
53. There are of course many pitfalls to this process, not least the fact that baptism dates are not an easy proxy for birth dates, that parish registers cover only a fraction of the population, and that highly mobile populations make family reconstitution impossible. These and other problems are discussed in the classic 'handbook' for demographers: E. A. Wrigley (ed.), *Identifying People in the Past* (London 1973).
54. See the online appendix to this volume at: www.bloomsbury.com/history-by-numbers.
55. The structure of the current (2010) British Standard Occupational Classification gives some idea of the difficulties: www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/index.html (accessed 4 December 2015).

INDEX

Page numbers in square brackets refer to endnotes; they are given after the page from which the referral is made (for example: concordance, 272 [328 n.13]).

- Access, 265, 271, 274
Acemoglu, Daron, 258
actuarial statistics, 25, 28
adverse selection, 255
age
 death, 180–182
 distributions, 69, 115, 190–193
 leaving home, 109–111
 marriage, 6–7, 11, 92–93, 214–216, 217, 223–224
age heaping, 295
age pyramids, 295
agency theory, 255, 295
aggregate data, 295
agricultural statistics, 34, 160, 196
algorithms, 269, 271, 279–280, 295
Allen, Robert C., 140–142, 144, 258
Anderson, Michael, 10, 206
Annals School, 2, 10
Annesley, James, 26, 28
anthropology, 252–253
anthropometric studies, 33, 154–155, 179–180, 281, 295
apprentice data, 190–193, 225, 268
ArcGIS, 273
archaeological data, 101–102, 273
archives, 275–277
arithmetic mean, 87–90, 92–93, 105–107
Arrow, Kenneth, 256
art market data, 191
Ashton, T. S., 140
asymmetric information, 255, 295
attribute data, 273
Austen, Jane, 281–282
autobiographies, 211, 281, 288
autocorrelation, 195, 196–203, 295
average man, 35, 112–113
average percentage growth rate, 148–149
averages, 87–95, 295

back projection, 243, 295
Bacon, Francis, 24
Bakhtin, Mikhail, 279
bar charts, 62–68, 295
Barker, Hannah, 280

Barnett, Vincent, 160, 196, 198–199
Bartley, Ken, 275
base period (base year), 130–133, 135–136, 138, 141–142, 295
basket of goods, 135, 139–140, 295
Bayes' theorem, 25 [312 n.14], 37, 37 [315 n.59]
Bayes, Thomas, 25
Becker, Gary, 255
bell-shaped distribution, 90, 112–113
benchmark years, 142, 246–247
better-documented samples, 212
Beynon, Huw, 206
bi-modal distribution, 91, 114, 116
bias, 40, 295
Bielefeld, J. F. von, 25
Big Data, 11, 277–278, 296
Bills of Mortality, 24
binomial distribution, 90, 112–113
biographies, 211
biometrics, 35–38, 296
biometric eugenics, 37
birth rates, 160–161
Black Death, 106–107
Black, William, 26, 28
Board of Trade, 28, 32
Bonaparte, Napoleon, 29
Boolean operators, 272 [328 n.12], 296
Borda ranking/count, 296
Botticini, M., 93–94, 223–224
Boyer, G. R., 191, 200, 220, 222–223, 250
Breusch–Gordon Lagrange multiplier test, 296
British statistical developments, 23–34
Broadberry, Stephen, 138
Buckle, Henry Thomas, 36
Bureau de Statistique, 29
Buringh, Eltjo, 176–177
Burke, Peter, 21
Burnette, Joyce, 188
business history, 7, 211–212

Cambridge Group for the History of Population and Social Structure, 11, 271, 277
Campbell, Bruce, 275
Carauna, Leonard, 197, 200

Index

- Carlos, Ann, 255
cartograms, 84–85, 101, 103, 273, 296
cases, 50, 205, 296
categorical data, 16–18, 45–47, 174, 296
causal variables, 78–81, 172
causality, 9–11, 19, 164–165, 176
cells, 51, 296
census enumerators' books, 10, 14, 206, 209, 215
[320 n.15], 216
census, of population
 British, 10–11, 17, 28, 32–33, 67
 collections of, 276–277
 use of, 104, 109–111, 206, 225, 243, 248, 268
central tendency, measures of, 87–95
ceteris paribus, 253, 259, 296
Chadwick, Edwin, 33
charts, 26, 73–82, 144–145, 296
chi-squared distribution, 36
chi-squared statistic, 167–169
Chicago School economists, 259
child labourers' data, 281
civil registration, 28, 215
Civil War (American), 248
Clapham, J. H., 140
class intervals, 57, 62, 65, 67, 296
class structure, 66
classification criteria, 16–18
cliometrics, 243, 245–246, 255, 296
cluster samples, 210–211, 296
coal mine data, 75–76
Cobb–Douglas production function, 296
Cobb, Richard, 3–4
Cochrane–Orcutt correction, 296
code and retrieve programs, 272
coding, 46, 271, 282, 289, 296
coefficient of determination, 184–185, 190, 296
coefficient of variation, 98–101, 297
Coleman, Donald, 129, 252
collocations, 272, 283
Colquhoun, Patrick, 25
comparability issues, 12
comparative advantage, 243 [323 n.32]
composite indices, 9, 130, 134–141, 297
computers and history, 3–6, 19, 263–291
 approaches, 284–286
 impact on historical work, 275–283
 research projects, 286–289
 software types, 263–275
 things to look out for, 289–290
concordance, 272 [328 n.13], 297
Conrad, A. H., 248–250
constant prices, 141, 297
consumer credit data, 255
consumer expenditure surveys, 34
consumption data, 131–132, 169–171, 176, 207
contingency coefficient, 164–169, 175, 297
contingency tables, 165–169, 174, 297
continuous data, 49, 165, 297
convergence (of economies), 99, 254
convict data, 154–155
 see also crime data
correlation, 9–10, 19, 36–37, 218–225, 297
correlation coefficient, 165, 174–176, 180, 297
cost-benefit analysis, 247–248, 297
cost of living indices, 34, 139–140, 142, 297
counterfactual history, 11–12, 244, 246–248,
 250–251, 297
court records, 212
covariance, 297
Crafts, N. F. R., 138, 200, 259
credit data, 255
crime data
 examples of, 55–58, 89, 91, 114, 116, 154–156,
 165–166, 185–188
 problems with, 17, 28
 use of, 207, 218–219
criminal trial data, 46, 207, 218–219, 288
cross sectional data, 163, 170, 297
cross-tabulations, 165
cumulative frequency distribution, 57–58, 297
curvilinear relationships, 173, 180
cycles, 153–154, 158, 160, 185–188, 297
data matrices, 50–52, 165, 287–288, 297
data processing, 297
data, types of, 45–50, 297
databanks, 276–277
databases, 265–271, 297
datasets, 17–18, 50, 298
Davenant, Charles, 24
Dbase 3, 265
death rates, 129–130, 160–161
deciles, 101–102, 298
deduction, 19 [310 n.45], 298
demographic evidence, 13, 33, 81, 154–155
demographic history, 7, 10–11, 24, 28, 177–178
demographic transition, 252 [324 n.41]
Dennison, Tracy, 191
denominator, 131 [317 n.3], 298
dependency ratio, 81
dependent variables, 78–79, 172, 176, 183–184,
 189–191, 199–200, 223–225, 298
descriptive statistics, 7–8, 45–84, 298
 bar charts, 62–68
 cartograms, 84–85
 data, questions about, 53–54
 data types, 45–50
 frequency distributions, 54–62
 graphs, independent and causal variables,
 78–82
 graphs, time series, 73–78
 histograms, 65–71

- pie charts and pyramid charts, 69–73
 regrouping data, 50–52
 tables and figures, 52–53
 word clouds, 83–84
- Desrosieres, Alain, 35
- detrended series, 80, 151–153, 157, 200, 298
- dichotomous variables, 46, 174, 298
- difference equations, 243
- difference-of-means test, 216–217
- digitization projects, 276
- discrete data, 49, 56, 165, 298
- dispersion, measures of, 95–111
 coefficient of variation, 98–101
 examples of, 104–111
 rank order, 101–104
 standard deviation and variance, 95–98
- distributions, 54–62, 87, 111–116, 298
- documentary sources, 211, 272, 276, 280–282
- Doda, Hilary, 53
- Dollar, D., 182–183
- Douglas, Mary, 252
- dowries, 223–224
- dummy variables, 174, 225, 298
- Dunstall, Graeme, 154, 156
- Dupin, Charles, 26
- Durbin-Watson test, 195, 197, 200
- Durkheim, Emile, 35
- econometric history, 11, 241–261, 298
 current, 254–256
 definition, 242–243
 economic growth, 257–259
 examples, 246–251
 history of, 244–246
 models, evidence and reality, 259–260
 theory, 251–253
- Econometric Society, 244
- Economic and Social Research Council (ESRC)
 data archive, 276, 288
- economic data, 132–133, 185–188
- education and family size data, 166
- EEC *see* European Economic Community
- elementary descriptive statistics, 45, 298
- employment statistics, 33, 80–81, 104, 188
see also unemployment statistics
- endogenous growth models, 254, 257
- Engerman, Stanley, 248
- entrepreneurship, 250
- equilibrium models, 245, 250, 252, 256
- error distribution, 112–113, 208, 212–213
- error theory, 8, 25, 35, 212–213, 302
- Espuelas, Sergio, 197, 201–202
- ESRC *see* Economic and Social Research Council
 data archive
- Etro, Federico, 191
- eugenics, 36–37
- European Economic Community (EEC), 51, 53–54, 298
- EViews, 265
- Excel, 264, 271
- exogenous variables, 257, 298
- expert systems, 271, 298
- exponential scales, 77–78, 298
- externality, 298
- extrapolation, 151, 184, 243, 260, 298
- factory workers, 206
- falsification, 19, 19 [310 n.45], 259, 278
- family and social relationships, 281–282
- family reconstitution, 81, 266, 268–271, 279–280, 287–288, 288 [330 n.53]
- Farr, William, 32–33
- Feinstein, Charles, 133–134
- female data, 80–81, 83, 104–105, 188, 215
- fields, data entry, 51, 265–268, 298
- figures, 52–53
see also descriptive statistics
- Filemaker Pro, 271
- film industry data, 78
- filters, 200, 298
- Fisher, Ronald, 36–37, 225
- Fisher's exact test, 37
- Fisher's Ideal Price Index, 299
- Fishlow, R., 247
- flatfile databases, 266, 299
- Floud, Roderick, 214–215, 217
- fluctuations, 144, 153–160, 299
- Fogel, R. W., 246–247, 248
- food prices data, 185–188, 200, 203
- Football League, 142–143
- footnotes for tables and figures, 52
- formalism, 244
- Foucault, Michel, 33
- Foxbase, 265
- France, 26, 29
- frequency, 54, 299
- frequency curves, 66–67, 70, 299
- frequency distributions, 54–62, 89, 91, 299
- frequency polygons, 66–67, 69–71, 299
- Furet, François, 2
- fuzzy searching, 84, 267, 272, 290
- Galton, Francis, 36–37, 113
- Gambling Act (1774), 25
- game theory, 256
- GDP *see* Gross Domestic Product
- Geary Kamis dollars, 259
- Geertz, Clifford, 18
- gender perspectives, 255
- General Register Office, 28, 32–33

Index

- Geographical Information Science (GISc), 273, 280–281
Geographical Information Systems (GIS), 84, 273–275, 299
geometric mean, 91–92
Giddens, Anthony, 40
Gini coefficient, 102, 108–109 [317 n.10], 299
GIS *see* Geographic Information Systems
GISc *see* Geographical Information Science
globalisation data, 99
GmapGIS, 273
GNP *see* Gross National Product
gold data, 158, 160
Gosset, William Sealy, 218
Gould, Stephen J., 18
government statistics, 23, 34
Graphical User Interface (GUI), 264
graphics packages, 264–265
graphs, 26, 74–82, 144–145, 299
Graunt, John, 24, 25
Great Divergence, 258
Gregory, Paul R., 99–100
Groningen, University of, 11, 258, 277
Gross Domestic Product (GDP), 12, 79–80, 99–100, 258–259
Gross National Product (GNP), 11, 34, 247
growth rates, 92, 146–149, 151, 257, 299
GUI *see* Graphical User Interface
Gupta, Bishnupriya, 225
Gutmann, Myron, 109–111
- Habakkuk, Sir John, 259
Halley, Edmund, 25
Harley, C. Knick, 138
Hatton, Tim, 191, 200, 220, 222–223
Hawke, G. R., 247
HDI *see* Human Development Index
hearth tax data, 57–59, 268
hedonic prices, 299
height distributions, 75–76, 90, 98, 101, 154–155, 179–180, 188–189
Heisenberg uncertainty principle, 16 [310 n.37]
HGIS *see* Historical Geographical Information Systems
Hill, Matthew J., 79
histograms, 62, 68–71, 299
histoire serielle, 2
histoire totale, 2
Historical Directories of England and Wales, 61
Historical Geographical Information Systems (HGIS), 273–275
historical relativism, 39
history from below, 280
Hobbes, Thomas, 24
- Hoffman, Walther, 138
l'homme moyen, 35, 112–113
Horrell, Sarah, 207, 218–219
household expenditure data, 46–47
household studies, 206
householding, 253
Hudson, Liam, 3
Hudson's Bay Company, 255
Human Development Index (HDI), 254, 261, 277
humanities GIS, 273
Humphries, Jane, 207, 218–219, 281
hypothesis, 176, 241, 248, 299
see also null hypothesis
hypothetico-deductive method, 19 [310 n.45]
- ICPSR *see* Inter-university Consortium for Political and Social Research
ideal-types, 112 [317 n.13]
import and export data, 51
income data, 47–48, 64–66, 68, 88, 170–171
independent random samples, 208–209, 212–213, 299
independent variables, 78–79, 172, 176, 183–184, 189–191, 299
index number problem, 137–138, 257, 299
index numbers, 130
indices, 9, 91, 129–144, 278, 299
composite, 134–141
formation of, 131–134
real, 141–144
induction, 38, 299
industrial output data, 17, 20, 133, 138–139, 181–182, 220–222
inferential statistics, 8, 37, 163–204, 299
autocorrelation, 195, 196–203
coefficient of determination, 184–185, 190
contingency coefficient, 164–169, 175
correlation coefficient, 165, 174–176, 180
dummy variables, 174, 225
examples, 185–193
lagged results, 177–180
multicollinearity, 195–202
multiple regression models, 4, 189–191
non-random error, 194–195
null hypothesis, 164–167
regression lines, 182–184
scatter graphs, 79, 169–173, 182
Spearman's rank correlation coefficient, 173, 180–182
inheritance data, 62–63
innovations, modelling of, 20, 78
Institute of Fiscal Studies, 93
institutions, 252, 255
instrumentalism, 259, 300
insurance data, 105–106, 158, 160
Intelligence Quotient (IQ), 18, 36

- Inter-university Consortium for Political and Social Research (ICPSR), 276–277
- Interactive Structured Query Language (ISQL), 268–269
- interquartile range, 101, 104, 300
- interval data, 49
- interviews, 206
- inverse probability, 37
- IQ *see* Intelligence Quotient
- Isherwood, Baron, 252
- Ishizu, Mina, 280
- ISQL *see* Interactive Structured Query Language
- Jackson, R. V., 138
- Judt, Tony, 3
- Kalman filter, 200
- King, Gregory, 14–15, 24, 47–48
- Labour Markets Database, 266
- lagged results, 177–180
- land ownership data, 168–169, 274–275
- land tax data
- examples of, 57–61, 70, 88–89, 90–91, 113–114
 - and software, 265–268, 287
- landholding inequality data, 101–103
- Laplace, Pierre Simon, 25
- Laspeyres index, 139, 300
- law of large numbers, 35
- Lazerev, Valery, 99–100
- Le Roy Ladurie, Emmanuel, 2, 286
- Lee, Clive, 251
- Lee, R. D., 243
- legal data, 95–97
- Lemire, Beverly, 105–106
- letters, sampled, 206
- Leunig, Tim, 188–189
- Leybourne, J., 200
- life tables, 24–25, 300
- line of best fit, 149, 183–184, 300
- see also* regression lines; trend lines
- linear trends, 146, 151, 157–158, 196, 243
- linguistic turn, 41
- see also* post-structuralism
- literacy data, 176–177
- Liverpool, University of, 276
- living standards, 17
- log linear analysis, 300
- logarithmic scales, 76–78, 79–80, 300
- logarithms, 300
- logistic growth curve, 300
- Logit analysis, 110–111, 300
- Logit models, 300
- London Lives (1690–1800), 280
- London statistical societies, 29, 32
- Lorenz curve, 81–82, 300
- lotteries, 208
- Lotus Approach, 271
- love and economic circumstances, 79–80
- Mackenzie, Donald A., 37
- macroeconomic, 300
- MAD (Manuals of Archival Description), 276
- Maddison, Angus, 277
- Manning, Patrick, 277
- market clearing, 252–253
- marriage age data, 6–7, 11, 92–93, 191
- marriage rate data, 11, 79–80, 225
- marriage statistics, 93–95, 223–224
- Marx, Karl, 35
- Marxist models, 243
- Mass Observation, 52, 54
- mathematical modelling, 11–12, 20
- matrices, 301
- see also* data matrices
- matrix notation, 51–52
- Max Planck Institute, 277
- McCloskey, Deidre, 241, 250
- McCulloch, J. R., 26
- mean *see* arithmetic mean
- mean increase per year, 146, 148
- mechanical objectivity, 40, 301
- media, printed, 7, 211, 276
- median
- definition, 90–91, 301
 - examples of, 88, 105–112
 - second quartile (Q2), 101, 111
- medical statistics, 11, 26, 28, 33, 34
- Meek, Ronald, 205
- Mendel, Gregor, 37 [315 n.56]
- merchants' letters, 206
- Metzer, J., 247
- Meyer, J. R., 248–250
- Michigan, University of, 276–277
- microeconomic, 301
- migration data, 84–85, 220, 222–223
- Mills, T. C., 200
- Milne, Joshua, 28
- Minitab, 265
- Minnesota, University of, 277
- Minns, Chris, 225
- modal class, 91, 301
- mode, 88, 91–93, 105–106, 301
- model building, 243, 251–253, 301
- modelling
- building of models, 243
 - criticism of, 260
 - general equilibrium, 245, 250, 252, 260
 - multivariate, 256 *see also* multiple regression models
 - neoclassical, 250, 251–253, 259

Index

- Moivre, Abraham de, 25
Mokyr, Joel, 178
monotonic relationships, 173, 180
moral hazard risks, 255
moral statistics, 33, 301
mortality data, 69, 71
Mosaic Data Archive project, 277
moving averages, 153–156, 301
Mulhall, Michael, 29–32
multicollinearity, 195–202, 301
multiple regression models, 4, 189–191
see also regression lines
multivariate analysis, 191, 301
multivariate modelling, 256
myopic choice, 256
- Napoleonic Wars, 29, 250
national accounts, 20, 34, 243, 277
see also industrial output data
National Historical Geographic Information System (NHGIS), 273
Nazi Party, 36
negative skew, 113–114, 301
negative trend, 145
neoclassical economics, 242–244, 251–253, 255, 301
New Economic History, 2–3, 245–246, 248, 250
new political history, 3
new urban history, 3
NHGIS *see* National Historical Geographic Information System
Nicholas, Steven, 255
Nietzsche, Friedrich, 32
nominal data, 45–47, 48, 50, 58, 91, 301
non-parametric statistics, 301
non-random error (in regression exercises), 194–195
normal distribution (or error curve), 112–113, 208, 212–213
normative theory, 301
see also eugenics; moral statistics; political arithmetic
null hypothesis, 164–167, 301
numerator, 131 [317 n.3], 301
numeric data, 45–50, 301
- objectivity, 38, 40–41, 245, 282–283, 301–302
occupational data, 71–72, 74
OCR *see* Optical Character Recognition
Odell, Kerry A., 158, 160
OECD *see* Organisation for Economic Co-operation and Development
Office for National Statistics, 264
Ogilvie, Sheilagh, 191
Old Bailey records, 46, 207, 218–219, 288
Olney, Martha J., 255
online data, 5
- Optical Character Recognition (OCR), 290, 302
oral history, 210–211
ordinal data, 45–48, 50, 53, 165, 302
Organisation for Economic Co-operation and Development (OECD), 99
outliers, 89, 302
Overton, Mark, 271
Oxley, Deborah, 188–189
- Paasche index, 139, 302
Pagani, Laura, 191
paintings data, 191
Paradox, 265
parametric statistics, 302
parish registers
 and computers, 268, 270
 examples of, 145, 147–148, 150–153, 214–215
 family reconstitution, 279, 287–288
 problems with, 243, 257
 use of, 11, 14, 260
Parkes, Frances, 83
parliamentary speeches data, 62–63
peace, meaning and use of, 281
Pearson, Egon, 36
Pearson, Karl, 36–40, 165, 174
percentages, 56–57, 61, 130 [317 n.2]
percentiles, 101, 302
Perrot, Michelle, 33
Petty, Sir William, 24
pictograms, 88, 91
pie charts, 71–74, 302
Pittsburgh, University of, 11, 277
Playfair, William, 26–27
Polanyi, Karl, 253
political arithmetic, 24, 302
Pomeranz, Kenneth, 258
poor relief system, 250
population, 205, 302
population data, 66, 68
see also census, of population
population growth, 11
Porter, G. R., 32
positive skew, 113–114, 302
positive trend, 145
positivism, 6, 38–40, 260, 302
post-modernism, 39
post-structuralism, 39, 279
Premium Bonds, 208
presentation of research, 278–279
price systems, 251
prices, 133–134, 139–140
printed media, 7, 211, 276
probability, 37, 164, 186, 188, 260, 302
see also significance tests
probability theory, 8, 25, 35, 212–213, 302
probate inventories, 206–207, 271

- problem-oriented approach, 284–285
 profit rates data, 75, 178
 property values, 180–182
 proxy figures, 15, 17, 140, 197, 257, 260, 302
 pub data, 51–52, 54
 pyramid charts, 71, 74, 302
- qualitative history, 39–40, 272
 quantitative history
 advantages of, 6–12
 in economic history, 2–3, 241–261
 history of, 1–3, 23–41
 limitations of, 12–21
 reaction against, 3–4, 39–41
 quartile deviation, 101, 104, 303
 quartiles, 101–104, 111, 303
 Quetelet, Adolphe, 35, 112–113
 quintiles, 101–104, 107–108, 303
- race, 36–37
 railways, 12, 170–172, 246–248
 random sample, 208–209, 212–213, 299
 range, 95, 303
 rank order, 90, 101–104, 180–181, 303
 ratio data, 49–50
 ratio values, 130
 rational choice, 255–256
 rationality, 20, 252, 254, 256, 260, 303
 RDBMS (relational database management system), 303
 real indices, 130, 141–144, 303
 real wages, 17, 99
 reciprocity, 253
 reconstitution of families and communities, 81, 266, 268–271, 279–280, 287–288, 288 [330 n.53]
 redistribution, 253
 references for tables and figures, 52
 regression, 9–10, 36, 182–193, 243, 303
 regression coefficient, 184, 186, 190–191, 199, 303
 regression lines, 182–184, 303
 see also multiple regression models
 relational databases, 266–269, 279, 303
 reliability
 of data, 13–16, 18–21, 28
 of statistical results, 18, 19, 37, 260
 representativeness, 21, 188, 208, 210, 212
 respectability ratio, 142, 144
 Retail Price Index, 139, 142
 Rickman, John, 28
 Robbins, Lionel, 253
 Rockoff, Hugh, 197, 200
 Rogers, Thorold, 34
 Rostow, W. W., 185–186
 Royal Africa Company, 255
 Royal Statistical Society, 35
- Sahlins, Marshall, 252–253
 sales data, 158–159
 sample mean distribution, 213–214
 samples, 10–11, 205–226, 303
 proportional samples, 209–210
 stratified samples, 210, 304
 surviving samples, 211, 304
 systematic samples, 206–207, 209–210, 304
 sampling error, 212–213, 303
 sampling theory, 3, 10, 209, 303
 SAS *see* Statistical Analysis System
 scatter graphs, 79, 169–173, 182, 303
 Schofield, Phillip R., 106–107
 Schofield, Roger, 14–15, 177, 243
 Schürer, Kevin, 10, 206
 scientific history, 6, 40–41
 scientific knowledge, 38–40
 Scottish Ministers' Widows Fund, 25
 Searle, Rebecca, 34
 seasonality, 80–81, 157–160, 182–183, 303
 semi-logarithmic graphs, 76–78, 303
 Senior, Nassau, 35
 serial history, 2, 303
 series of first differences, 196–197, 303
 significance
 statistical, 36, 37 [315 n.58], 186, 217–225, 304
 statistical vs historical, 19, 37–38, 164, 278
 significance tests, 37, 199, 217–224, 304
 Silberling man, 140
 Silberling, N. J., 140
 Simpson, Thomas, 25
 Sinclair, Sir John, 26
 skewed distributions, 92–93, 95, 113–114, 304
 slave data, 66, 69, 248–250, 280
 slope coefficient, 184, 186, 190–191, 199, 303
 slopsellers data, 105–106
 smallpox data, 10, 75–76, 188–189
 Smith, Julia, 168–169
 smoothing data, 154–156
 Sneath, Ken, 207, 218–219
 Snell, K. D. M., 80–81
 social capability, 254
 social reform, 29, 32
 social savings, 244, 246–247, 304
 software for historians, 263–275
 Sokoloff, Kenneth L., 182–183
 Solow, R., 246
 source-oriented approach, 285–286
 sources for tables and figures, 52
 Soviet commodity markets, 160, 196, 198–199
 spatial and mapping software, 273–275
 spatial data, 273
 Spearman's rank correlation coefficient, 173, 180–182
 Spencer, Herbert, 241
 spreadsheets, 102, 264

Index

- SPSS *see* Statistical Package for the Social Sciences
SQL *see* Structured Query Language
standard deviation, 95–98, 106–107, 112, 213–214, 304
standard scores, 97–98
standardization of data, 271
Stata, 265
Statistical Analysis System (SAS), 265
Statistical Package for the Social Sciences (SPSS), 264–265
statistics
 definition of, 25–26, 304
 as language, xvi, 12, 39, 290
 techniques, choice of, 18
Stone, Gilbert, 26, 28
Stone, Lawrence, 3
story plot data, 62, 64
Stouffer, Samuel, 4
stratified samples, 210, 304
strike data, 131
structured data, 304
Structured Query Language (SQL), 268–269, 304
student's *t*-test, 218 [320 n.16]
 see also t-test
stylometric analysis, 272, 288
subjectivity, 38, 40–41, 245, 282–283, 301–302
substantivism, 253, 304
suicide, 35–36
sumptuary laws data, 53
Sun Life Assurance Society, 28
survey, 304
surviving samples, 10, 211, 304
systematic samples, 209–210, 304
- t*-statistics, 188, 191, 220, 223, 225, 305
t-test, 218, 218 [320 n.16], 225, 305
tables, 50–53
TACT (Text-Analysis Computing Tools), 272
tax data, 14–15, 107–108
technological change, 258
Temin, Peter, 245
textile data, 48–49, 61–62, 73–75, 178–179, 220–222
textual analysis software, 7, 83–84, 271–272, 285, 304
textual data, 304
textual processing, 271–272, 281, 288
thick description, 283 [330 n.50]
Thompson, Paul, 10, 210
Tilly, Charles, 4
time inconsistency, 256
time series, 129–161, 304
 and Big Data, 278
 and causal analysis, 9–10
 fluctuations, 153–160
 graphs of, 74–78
- indices, 130–144
influences, 144–145
scatter graphs, 169–171
trends, 145–153
 vital statistics, 160–161
titles for tables and figures, 52
township data, 266–268
trade statistics, 13, 51, 61, 69, 72
trend lines, 149–153, 183, 304
trends, 145–153, 304
tri-modal distribution, 91, 116
tungsten ore supply data, 197, 200
twentieth-century statistics, 33–35
- unemployment statistics, 13–15
 see also employment statistics
University of Minnesota Population Centre, 277
- Van Zanden, Jan Luiten, 176–177
variables, 50, 305
variance, 95–96, 305
Variance Inflation Factor (VIF), 197, 199, 305
vectors, 51, 53, 305
vehicle allocation data, 99–100
Victorian statistical movement, 29–33
VIF *see* Variance Inflation Factor
virtual history, 244, 244 [322 n.14], 305
A Vision of Britain through Time, 273, 280
vital events, 11, 154, 243, 287
 see also vital statistics
vital statistics, 160–161, 260
voluntarism, 29–33
Voth, Hans-Joachim, 188–189
- wage data, 104, 135–138, 140–142, 144, 182–183, 188, 196–197, 220, 222
wage estimates, 17
Wallace, Robert, 25
Wallis, Patrick, 225
war and statistics, 34, 80
wealth inequality, 82, 93–94
Weatherill, Lorna, 207
Webb, Cliff, 225
Weber, Max, 112 [317 n.13]
Webster, Alexander, 25
Weidenmier, Marc D., 158, 160
weights, 20, 134–139, 305
weights and measures, 25
welfare spending, 197, 201–202
Welfare State, 34, 107–108
Westminster Historical Database, 280
Williamson, Jeffrey G., 99, 250
wills, 212
Withington, Phil, 281
women's data, 80–81, 83, 104–105, 188, 215
Woppard, Matthew, 10, 206

- word clouds, 82–84, 305
word processing, 278–279
Wordcruncher, 272
WordSmith tools, 272
workhouse data, 158–160
Wrigley, E. A., 14–15, 92–93, 177, 243

x-axis, 305
y-axis, 77, 305
Yule, G. Udney, 37, 165
Z scores, 97–98, 305

