

Register-based Statistics

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

Register-based Statistics

Registers and the National Statistical System

Third Edition

Anders Wallgren and Britt Wallgren

*Formerly at the Department of Research and Development at
Statistics Sweden*

WILEY

This third edition first published 2022
© 2022 John Wiley & Sons Ltd

Edition History

John Wiley & Sons Ltd (2e, 2014)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Anders Wallgren and Britt Wallgren to be identified as the authors of this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

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

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting scientific method, diagnosis, or treatment by physicians for any particular patient. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

A catalogue record for this book is available from the Library of Congress

Hardback ISBN: 9781119632375; ePub ISBN: 9781119632689; ePDF ISBN: 9781119632641;

Obook ISBN: 9781119632672.

Cover images: Courtesy of Anders Wallgren and Britt Wallgren

Cover design by Wiley

Set in 11/12.5pt TimesNewRomanMTStd by Integra Software Services Pvt. Ltd, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Index to Charts		xi
Preface		xix
Chapter 1	Censuses, Sample Surveys and Register Surveys	1
1.1	The national statistical system	2
1.2	The traditional census-based system	3
1.3	New sources: Administrative registers and Big data	5
1.4	Basic concepts and terms	7
	1.4.1 <i>What is a register?</i>	7
	1.4.2 <i>Databases, records and observations</i>	8
	1.4.3 <i>What is a register survey?</i>	10
	1.4.4 <i>A register survey: The Income and Taxation Register</i>	12
1.5	New demands and opportunities require new methods	14
	1.5.1 <i>A new paradigm is necessary</i>	14
	1.5.2 <i>New statistical methods</i>	15
	1.5.3 <i>The basic principles of register-based statistics</i>	18
1.6	Preconditions for register-based statistics	19
	1.6.1 <i>Reliable administrative systems</i>	20
	1.6.2 <i>Legal base and public approval</i>	21
	1.6.3 <i>Political support to strengthen the statistical system</i>	23
Chapter 2	The Transition to a Register-based Production System	25
2.1	First obstacle: How to gain access to microdata?	26
2.2	Protection of privacy and confidentiality	26
2.3	Second obstacle: How to take care of dirty data?	29
2.4	The new production process	30
	2.4.1 <i>Contacts with administrative authorities</i>	31
	2.4.2 <i>Metadata have a new role</i>	31
	2.4.3 <i>Anonymisation of identity numbers</i>	32
	2.4.4 <i>Editing of a single administrative register</i>	33
	2.4.5 <i>Organising the work with administrative registers</i>	36
2.5	Third obstacle: The national registration system	37
	2.5.1 <i>Legislation governs access to data</i>	38
	2.5.2 <i>Too many registers, but no good registers – what to do?</i>	38
	2.5.3 <i>Legislation rules obligations to report and what to report</i>	40
2.6	Why has the census been so important?	41

2.7	Creating the register system	42
	2.7.1 <i>Where do you live?</i>	43
	2.7.2 <i>Where do you work?</i>	45
	2.7.3 <i>With whom do you live?</i>	46
	2.7.4 <i>A centralised or decentralised national system?</i>	47
2.8	Register surveys and estimation methods	48
2.9	A traditional census or a register-based census?	49
Chapter 3	The Nature of Administrative Data	51
3.1	Comparing questionnaire and register data	51
	3.1.1 <i>A questionnaire to persons compared with register data</i>	51
	3.1.2 <i>An enterprise questionnaire compared with register data</i>	54
3.2	Enterprise registers for combined use	56
	3.2.1 <i>Corrections in accounting data</i>	57
	3.2.2 <i>Missing values in accounting data</i>	58
	3.2.3 <i>Administrative and statistical information systems</i>	59
3.3	Measurement errors in questionnaire and register data	60
	3.3.1 <i>Measurement errors</i>	61
	3.3.2 <i>Taxation errors</i>	62
Chapter 4	Building the System – Record Linkage	65
4.1	Record linkage	65
4.2	Record linkage in the Nordic countries	66
4.3	Deterministic record linkage	68
4.4	Creating variables by adjoining and aggregation	70
4.5	Probabilistic record linkage	73
4.6	Four causes of matching errors	79
4.7	The statistical system and record linkage	82
Chapter 5	Building the System – Quality Assessment	85
5.1	Four quality concepts	85
5.2	Making an inventory of potential sources	87
5.3	How can a source be used?	87
5.4	Quality assessment in a register-based production system	90
	5.4.1 <i>Analysing metadata</i>	91
	5.4.2 <i>Analysis and data editing of the source</i>	92
	5.4.3 <i>Comparing a source with the base register</i>	92
	5.4.4 <i>Comparing a source with surveys with similar variables</i>	93
5.5	Output data quality and quality of estimates	94
	5.5.1 <i>Analysing quality with a test census</i>	94
	5.5.2 <i>Analysing quality with samples from the new register</i>	95
	5.5.3 <i>Analysing quality with area samples</i>	96
	5.5.4 <i>Measuring quality of basic register variables with the LFS</i>	98
5.6	A coordinated system of registers	98
	5.6.1 <i>Are the base registers a coordinated system?</i>	98
	5.6.2 <i>Quality indicators at the system level</i>	99
5.7	Using the quality indicators	101

Chapter 6	Building the System – Editing Register Data	107
6.1	Editing in register surveys	108
6.2	Editing of a single administrative register	109
6.3	Consistency editing	110
	6.3.1 <i>Consistency editing – is the population correct?</i>	111
	6.3.2 <i>Consistency editing – are the units correct?</i>	118
	6.3.3 <i>Consistency editing – are the variables correct?</i>	120
6.4	Case studies – editing register data	121
	6.4.1 <i>Editing work within the Income and Taxation Register</i>	121
	6.4.2 <i>Editing work within the Income Statement Register</i>	123
	6.4.3 <i>What more can be learned from these examples?</i>	124
6.5	Editing, quality assessment and survey design	125
	6.5.1 <i>Survey design in a register-based production system</i>	125
	6.5.2 <i>Survey design – management problems</i>	127
	6.5.3 <i>Total survey error in a register-based system</i>	128
Chapter 7	Building the System – The Population Register	129
7.1	Inventory of sources	131
	7.1.1 <i>Time references</i>	131
	7.1.2 <i>Activities or ‘signs of life’</i>	131
7.2	The Population Register based on full information	133
	7.2.1 <i>Object types – Changing and unchanging registers</i>	133
	7.2.2 <i>Variables with different functions in the system</i>	134
	7.2.3 <i>Updating the Population Register</i>	136
	7.2.4 <i>Registers and time</i>	137
	7.2.5 <i>Variables and time</i>	140
7.3	The Population Register in new register countries	140
	7.3.1 <i>Different systems of identity numbers</i>	141
	7.3.2 <i>Problems in countries without a central Population Register</i>	142
	7.3.3 <i>How to improve coverage of the Population Register</i>	143
	7.3.4 <i>Inventory of sources – addresses and time references</i>	146
7.4	Methods to measure and improve quality	148
	7.4.1 <i>Three kinds of surveys should be combined</i>	148
	7.4.2 <i>A new register-based system for statistics on persons</i>	150
7.5	Conclusions	151
7.6	Challenges in old register countries	152
Chapter 8	The Population Register – Estimation Methods	155
8.1	Estimation in sample surveys and register surveys	156
	8.1.1 <i>Estimation methods for register surveys that use weights</i>	157
	8.1.2 <i>Calibration of weights in register surveys</i>	157
8.2	Calibration of weights – the Swedish LFS	161
	8.2.1 <i>Use of auxiliary information in the LFS</i>	161
	8.2.2 <i>Nonresponse bias in the LFS</i>	162
8.3	Calibration – where do people live?	163
8.4	Methods to handle overcoverage	167

Chapter 9	Defining Register Populations – Coverage Errors	171
9.1	Defining a register's object set	172
9.1.1	<i>Defining a population</i>	172
9.1.2	<i>Can you alter data from the National Tax Agency?</i>	176
9.1.3	<i>Defining a population – the Farm Register</i>	176
9.1.4	<i>Defining a population – integrated registers</i>	178
9.2	Defining a calendar year population	179
9.2.1	<i>Defining a population – frame or register population?</i>	180
9.2.2	<i>Sampling paradigm versus register paradigm</i>	184
Chapter 10	Building the System – The Business Register	185
10.1	The Business Register and the National Accounts	185
10.2	The base register for economic statistics	187
10.3	The scope of the register and choice of object types	188
10.3.1	<i>The register with legal units and local units</i>	189
10.3.2	<i>The register with enterprise units and kind of activity units</i>	191
10.4	Inventory of sources	195
10.5	Creating and maintaining the Business Register	198
Chapter 11	The Business Register – Estimation Methods	201
11.1	Multi-valued variables	202
11.2	Estimation methods	205
11.2.1	<i>Occupation in the Activity and Occupation Registers</i>	206
11.2.2	<i>Industrial classification in the Business Register</i>	210
11.2.3	<i>Estimates from different register versions</i>	213
11.3	Application of the method	214
11.3.1	<i>Change of industry and time series quality</i>	215
11.3.2	<i>Transformation of weights</i>	217
11.4	A decentralised or centralised statistical system?	218
11.4.1	<i>The Calendar Year Register and the National Accounts</i>	219
11.4.2	<i>Choosing the best source for the National Accounts</i>	220
11.5	Conclusions	224
Chapter 12	Censuses, Sample Surveys and Register Surveys – Conclusions	227
12.1	Attitudes towards the register-based census	227
12.2	The new national statistical system	231
12.2.1	<i>The system of base registers</i>	232
12.2.2	<i>Activity registers and longitudinal registers</i>	234
12.3	Survey design	237
12.3.1	<i>Sample survey design</i>	237
12.3.2	<i>Register survey design</i>	238
12.3.3	<i>Creating register variables</i>	241
12.4	Survey quality	245
12.4.1	<i>Quality of registers and register surveys</i>	246
12.4.2	<i>The integration process – integration errors</i>	247
12.4.3	<i>Frame errors</i>	247

12.5	Organising the new production system	248
	12.5.1 <i>Enterprise architecture and the register system</i>	248
	12.5.2 <i>The register system and data warehousing</i>	249
	12.5.3 <i>Missing values – a system-based approach</i>	252
12.6	Final remarks	254
	12.6.1 <i>The Statistical Population Register</i>	254
	12.6.2 <i>The system of base registers</i>	255
	References	257
	Index	261

Index to Charts

Chapter 1	Censuses, Sample Surveys and Register Surveys	1
Chart 1.1	The Statistical System in a country can consist of the following actors	2
Chart 1.2	The Statistical System in a country can consist of the following surveys	3
Chart 1.3	Employment with census data	4
Chart 1.4	Employment with register data	4
Chart 1.5	The year of establishing new statistical registers in the Nordic countries	5
Chart 1.6	Example of a register and data matrix	9
Chart 1.7	A conceptual database model with three database tables	9
Chart 1.8	A database on individuals with three database tables	10
Chart 1.9	Two data matrices for different statistical purposes	10
Chart 1.10	From administrative registers to statistical registers – overview	11
Chart 1.11	From administrative registers to statistical registers – register processing	11
Chart 1.12	Different data sources for the Income and Taxation Register (I &T)	13
Chart 1.13	Employees by economic activity, November 2004, thousands	17
Chart 1.14	Register-based census statistics for one small municipality in Sweden	18
Chart 1.15	Four principles for using administrative registers for statistics	19
Chart 1.16	Two preconditions for using administrative registers for statistics	20

Chapter 2	The Transition to a Register-based Production System	25
Chart 2.1	The transition entails a fundamental change of statistical methods	25
Chart 2.2	An administrative register – unprocessed data in the input database	27
Chart 2.3	Corresponding statistical register – processed data in throughput database	28
Chart 2.4	Transforming administrative registers into statistical registers	30
Chart 2.5	Metadata system for input data from administrative registers	32
Chart 2.6	Identity database	33
Chart 2.7	Errors in profit and loss statements from limited companies, SEK million	35
Chart 2.8	Profit and loss statements for limited companies, SEK million	35
Chart 2.9	Registration of persons & demographic events in a Latin American country	38
Chart 2.10	Uncoordinated work with registers	39
Chart 2.11	Cooperation regarding the central population register	40
Chart 2.12	Sweden's system 1967–1984	43
Chart 2.13	Sweden's system 1985–2010	45
Chart 2.14	Sweden's system 2011	46
Chart 2.15	Survey design: reduce costs and/or improve quality	49
Chapter 3	The Nature of Administrative Data	51
Chart 3.1	Yearly turnover for the same enterprises in three sources, USD million	54
Chart 3.2	Distance for each ordered observation	55
Chart 3.3	Complete groups of enterprises	55
Chart 3.4	Three registers, Invoice Register, Client Register, Item Register	56
Chart 3.5	Statistical Sales Register for January 2012 – four transactions	57
Chart 3.6	Administrative Invoice Register	58
Chart 3.7	Statistical Invoice Register	58
Chart 3.8	Administrative Item Register	58
Chart 3.9	Statistical Item Register	58
Chart 3.10	Measurement errors – comparison of data collection methods	62

Chapter 4	Building the System – Record Linkage	65
Chart 4.1a	Matching without errors in the matching key (PIN)	68
Chart 4.1b	Matching with errors in the matching key (PIN)	68
Chart 4.1c	Matching with errors in the matching key (First name and Surname)	69
Chart 4.1d	Matching without errors in the matching key (PIN)	69
Chart 4.2a	The relations between persons, activities and establishments	71
Chart 4.2b	Wage sums for persons and establishments created by aggregation	72
Chart 4.2c	Industry and sex as derived variables for jobs created by adjoining	72
Chart 4.2d	Industry and number of employees as derived variables	73
Chart 4.3	Spelling errors, spelling variations and unstandardized addresses	74
Chart 4.4	Two ways of searching for duplicates in a Birth Register	76
Chart 4.5	Calculation of probabilities for outcomes for 1st name, 1st and 2nd surname	77
Chart 4.6	The eight outcomes for the matching key: 1st name, 1st and 2nd surname	78
Chart 4.7	Some outcomes for the matching key: Name and date of birth of the mother	79
Chart 4.8	Yearly turnover for the same legal units in three sources, USD million	80
Chart 4.9	Comparing gross yearly pay in quarterly and annual registers	81
Chapter 5	Building the System – Quality Assessment	85
Chart 5.1	A register survey: From administrative registers to statistical estimates	85
Chart 5.2	Statistical registers that are suitable for early development	87
Chart 5.3	Input and output data and the production process	89
Chart 5.4	The work with quality assessment of an administrative source	90
Chart 5.5	Indicators A1-A9 of input data quality – Relevance	91
Chart 5.6	Indicators B1-B7 of input data quality – Accuracy	92
Chart 5.7	Indicators C1-C5 of accuracy when comparing with the base register	93
Chart 5.8	Indicators D1-D4 of accuracy when comparing with related surveys	94
Chart 5.9	Coverage errors in the population register for Galapagos 2015	95

Chart 5.10	Coverage errors and area sample estimates of the population by district	97
Chart 5.11	Creating registers by linking different kinds of statistical units	99
Chart 5.12	Indicators E1-E4 of the quality of base registers in the system	100
Chart 5.13	Information from the administrative authority – relevance	101
Chart 5.14	Information from analysis and data editing of the source – accuracy	102
Chart 5.15	Information from integrating the source with base registers – accuracy	103
Chart 5.16	Information from integrating the source with related surveys – accuracy	104
Chapter 6	Building the System – Editing Register Data	107
Chart 6.1	Editing in sample surveys and register surveys	108
Chart 6.2	Automatic editing and imputation. Tax returns, 464 567 small enterprises	110
Chart 6.3a	The Business Register and the Annual Pay Register	112
Chart 6.3b	After matching the Business Register and the Annual Pay Register	113
Chart 6.3c	The statistical Annual Pay Register	114
Chart 6.4a	Combining the Business, Annual and Quarterly Pay Registers	115
Chart 6.4b	Final version of the statistical Annual Pay Register	116
Chart 6.5	All relevant sources should be considered simultaneously	117
Chart 6.6	Comparing gross annual pay in a quarterly and yearly source, microdata	119
Chart 6.7	The unit problem in business data	120
Chart 6.8	Comparing gross annual pay in QGP and AGP, microdata	120
Chart 6.9	Comparing gross annual pay in QGP and AGP, macrodata	121
Chart 6.10	The system of registers and surveys that was analysed	127
Chapter 7	Building the System – The Population Register	129
Chart 7.1	Decentralised but coordinated process to create registers on persons	130
Chart 7.2	The system of registers that constitutes the Population Register at the NSO	134
Chart 7.3	Different types of variables in the Population Register	135
Chart 7.4a	The Population Register at December 31, 2012	136

Chart 7.4b	Notifications regarding demographic events delivered 1 February 2013	136
Chart 7.4c	Old register matched with the new notifications	137
Chart 7.4d	Updated register 1 February 2013	137
Chart 7.4e	Three versions	137
Chart 7.5	Calendar year register for 2012	139
Chart 7.6	Events register for 2012 regarding change of address	139
Chart 7.7	Historical register regarding change of address	139
Chart 7.8	Longitudinal register for 2010–2012	139
Chart 7.9	Many registers, but no register that covers the entire population	142
Chart 7.10	The first steps towards a new statistical system	147
Chart 7.11	Different parts of the target population and the desired variables	148
Chart 7.12	Estimation of undercoverage	149
Chart 7.13	Coverage errors and area sample estimates of the population by district	150
Chart 7.14	Coverage errors and register-based sample estimates of population	151
Chart 7.15	A register survey: From administrative registers to statistical estimates	151
Chapter 8	The Population Register – Estimation Methods	155
Chart 8.1	Register of all persons in two small districts	157
Chart 8.2a	Employed persons by education and industry	158
Chart 8.2b	Employed persons by education and industry, missing values are shown	158
Chart 8.3	Persons by Education and Industry, adjusted for missing values	161
Chart 8.4	Nonresponse rates in the Swedish LFS	161
Chart 8.5	Nonresponse bias in the LFS, December 2011–2015. Relative bias, percent	162
Chart 8.6	Population estimates, Exercise 3	165
Chart 8.7	Exercises 1-5: Reducing coverage errors in population estimates	166
Chart 8.8	Estimates for Exercise 1 ('RC') and Exercise 5	166
Chapter 9	Defining Register Populations – Coverage Errors	171
Chart 9.1	Methodological differences between social and economic statistics	171
Chart 9.2	Frame populations are created before register populations	173
Chart 9.3	Undercoverage in an administrative register	177

Chart 9.4	Object sets when matching two registers	178
Chart 9.5	Calendar year register for the population in a (small) municipality	179
Chart 9.6	Calendar year register for 2013 for enterprises in a particular (small) region	180
Chart 9.7	The November frame and the Calendar Year Register 2004	181
Chart 9.8	Over- and undercoverage as number of legal units in the November frame	182
Chart 9.9	Errors due to undercoverage in the November frame 2004, SEK million	182
Chart 9.10	Undercoverage in November frame 2004, non-financial sector by industry	183
Chart 9.11	Overcoverage in SBS based on November frame 2004, by industry	183
Chapter 10	Building the System – The Business Register	185
Chart 10.1	The system of economic statistics	185
Chart 10.2	Inconsistent sources with wage sums in the National Accounts	187
Chart 10.3	Legal units by institutional sector and industry	190
Chart 10.4	Industry improved	190
Chart 10.5	Many economic activities	190
Chart 10.6	Different types of units in the Business Register	191
Chart 10.7	Legal units and enterprise units	192
Chart 10.8	Yearly turnover for the same legal units in three sources, USD million	192
Chart 10.9	Comparing gross yearly pay in quarterly and annual registers	193
Chart 10.10a	Each administrative system has its own object set – the effect on microdata	193
Chart 10.10b	Each administrative system has its own object set – different combinations	194
Chart 10.11	Sources with information for the Business Register	197
Chapter 11	The Business Register – Estimation Methods	201
Chart 11.1	Employees by industry, November 2004, thousands	202
Chart 11.2	Industry and number of employees as derived variables	203
Chart 11.3a	Calendar year register for the population of persons during 2013	204
Chart 11.3b	Average population 2013	204
Chart 11.4	Number of employed and wage sums in different registers	204

Chart 11.5	Register 1 continued – persons and weights	205
Chart 11.6	Employed by industry	205
Chart 11.7	Job Register with occupational data for six persons	206
Chart 11.8	Traditional register on persons with occupational information	207
Chart 11.9	Employed persons by occupation, traditional alternative 1	207
Chart 11.10	Register of combination objects: person • occupation	208
Chart 11.11	Employed persons by occupation according to two alternatives	208
Chart 11.12	Register on jobs of persons with occupational data	210
Chart 11.13	Persons and full-time employed by occupation, three alternatives	210
Chart 11.14a	Business Register year 1: Data matrix for establishments (local units)	211
Chart 11.14b	Business Register year 2: Data matrix for local units	211
Chart 11.14c	Number of employees by industry, traditional estimates	211
Chart 11.15	Data matrix with combination objects: local unit • industry	212
Chart 11.16	Number of employees by industry, estimated with combination objects	212
Chart 11.17	Four different registers belonging to the Business Register	213
Chart 11.18	Number of employees by industry	214
Chart 11.19	Employees by industry November 2004, thousands	215
Chart 11.20	Estimation of turnover within one industry using two methods	216
Chart 11.21	Turnover in an industry, two estimates	217
Chart 11.22	Transformation of weights	218
Chart 11.23	Calendar year register for 2007, legal units by sector and industry	219
Chart 11.24	Coverage errors in the Business Register compared with the Farm Register	220
Chart 11.25	Statistics Sweden's present way of using business surveys	222
Chart 11.26	A new estimation method based on integrated microdata	223
Chapter 12	Censuses, Sample Surveys and Register Surveys – Conclusions	227
Chart 12.1	Monthly population data for two Swedish municipalities 2000–2020	231

Chart 12.2	A register-based statistical system	232
Chart 12.3	Transitions from 2003 to 2004, men 25–54 years. % of each category 2003	236
Chart 12.4	Per cent employed after completing education 1987–1992	237
Chart 12.5	The three parallel processes with a register survey	240
Chart 12.6	Classification errors in the old and new Employment Register 1993	244
Chart 12.7	Model errors in the Employment Register 1993	244
Chart 12.8	Quality of estimates for two domains in the Employment Register	245
Chart 12.9	A longitudinal income register, data for six persons	246
Chart 12.10	Different parts of the survey production process	249

Preface

Survey sampling and register surveys – what is the main difference?

What is the main difference between a book on survey sampling and a book on register-based statistics? Both are books on survey methodology, but there is one important difference that should be understood from the outset:

- Books on survey sampling discuss *one* sample survey. We have *one* population and collect data for *one* survey. The sampling books then discuss how this sample survey can be designed in different ways.
- Instead, books on register surveys must have a systems approach. When discussing one register survey, we must also understand the role other registers play in the system of registers used when the statistical register in question is created and evaluated.

During the 1960s, Svein Nordbotten (1967) at Statistics Norway developed ideas on statistical information systems and explained that administrative sources should be used for statistical purposes. The *statistical information system* concept is quite different from the *statistical survey* concept, and the former is suitable for organisations that regularly collect data and produce statistics. He subsequently introduced what he called *statistical file systems*, or what we now call register systems used for the production of official statistics.

The register system is not just a set of registers. The system also has methodological implications regarding how it should be designed, coordinated and used. This was made clear in Statistics Denmark's book (Danish version 1994, English translation 1995). It explains how a system of statistical registers should be designed to produce a register-based population and housing census. The Danish book was the starting point for our work with registers at Statistics Sweden. The book is discussed in Sections 1.5.2 and 12.1.

Official statistics or corporate statistics?

The National Statistical Office is usually the largest organisation that collects data and produces official statistics in a country. However, statistical information systems are also important in other kinds of organisations. *Business intelligence* is a term that is often used for statistical information systems within corporations.

Bo Sundgren from Statistics Sweden was visiting professor at Linköping University during 1984–1986. He started a research programme on statistical information systems. At that time, we taught statistics to students of Business Administration and at the statistics programme in Linköping. We started our own research project ‘Corporate information systems – Statistical analysis with the enterprise’s administrative data’. We had contacts in this project with several companies, and we tutored many students who worked with papers for their degree. We recommend that university statisticians try something similar – this can be a good way to start statistical research on how to work with administrative data. Section 3.2 contains a short description of this area. If we had stayed in Linköping, perhaps we would have written books on corporate statistics. Instead, we went to Statistics Sweden, resulting in this book that is devoted to the national statistical system and how registers should be used to improve this system.

How did we work with this book?

Teaching, teaching, and teaching – this has been an integral part of our work with this book. We started with study circles for the teams working with different registers at Statistics Sweden. Together, we analysed the registers and discussed methodological problems.

We have continued along these lines since leaving Statistics Sweden – combining teaching and discussions with colleagues. We have visited some European countries as well as several countries in Latin America and the Caribbean, where our work was supported by the Inter-American Development Bank.

This approach has been very stimulating and has given us a broad picture of just where the important problems are. We have also learned how the subject area should be described and explained to statisticians in countries that are starting work with registers.

We hope that *Register-based Statistics – Registers and the National Statistical System* and its proposals will stimulate the discussion of statistical registers and register systems and provide support to those working with register surveys at national statistical offices.

Örebro, Sweden

Anders Wallgren
Britt Wallgren
ba.statistik@telia.com

CHAPTER 1

Censuses, Sample Surveys and Register Surveys

National statistical offices use three kinds of survey methodology when producing official statistics based on microdata: methods for *censuses*, for *sample surveys* and for *register surveys*. This book deals with the third kind of methodology – methods for register surveys, where instead of collecting data through interviewers and questionnaires, *administrative registers* from different administrative systems are adapted and processed to create *statistical registers* that are used to produce the desired estimates.

We introduce several concepts and principles that should be used when discussing register surveys. These concepts and principles form the methodological bases for this kind of survey. There is a growing interest in this area. Many countries increasingly use administrative registers for statistical purposes, and there is a growing demand for an understanding of register survey methodology.

However, preconditions differ – in some countries the preconditions are good, while in other countries there can be obstacles that make it difficult to use data from some administrative systems. We discuss such obstacles and how the national statistical system can be improved to reduce the problems. We give special attention to countries desiring to take the first steps towards a register-based statistical system.

The statistical offices in the Nordic countries started using registers during the 1960s; and experiences from these countries are important in understanding how statistical systems in other countries could be improved.

Purpose of this book

Our purpose is to describe and explain the methods that should be used for register surveys. Conducting a register survey means that a new *statistical register* for a specific subject matter is created with existing sources. The statistical register is then used to produce estimates required for the survey. What methods should be used in creating such a statistical register? One or more administrative registers are used when a new statistical register is created, and the statistical register can differ from the administrative sources in many ways.

A *system of statistical registers* consists of a number of registers that can be linked to each other. In the Nordic countries, the national statistical offices have developed systems of registers that are used in the production of statistics. When new statistical registers are created, this register system becomes an important source that can be used together with different administrative sources. Another purpose of the book is to explain how such register systems should be designed and used in the production of statistics.

When a national statistical office starts using more administrative sources, the *statistical production system* of that office gradually changes. From a system based on enumerators or interviewers, address lists and maps, the system will become increasingly register based. Sample surveys will be based on the Population Register or the Business Register – variables in sample surveys can come from administrative registers as well as from telephone interviews or questionnaires. In addition to the change in methods used for sample surveys, new kinds of register-based statistics can also be produced. A third purpose of the book is to explain how administrative registers can be used to change the statistical production system of a national statistical office to improve cost efficiency and statistical quality.

1.1 The national statistical system

Official statistics in a country is produced by the national statistical system. We use two different interpretations of this system:

- The *system of actors* that is responsible for the official statistics.
- The *system of surveys* (censuses, sample surveys and register surveys) that these actors carry out to generate the desired microdata and estimates.

The actors in the system should cooperate to avoid duplicate work and conflicting or inconsistent surveys. The development of new register surveys requires that the national statistical office gains access to administrative registers that have been created and maintained by ministries or other administrative authorities. Data sharing and cooperation will then become necessary.

Chart 1.1 The Statistical System in a country can consist of the following actors:

1. The National Statistical Office¹ (NSO)
2. The National Advisory Council on Statistics
3. The Interagency Coordinating Committee on Statistics
4. The statistical offices of the Ministries
5. Statistical bodies of Regional Governments
6. Statistical bodies of Municipalities
7. Statistical bodies of public authorities

These actors interact with the political level that decides on legislation and funding of the national statistical system.

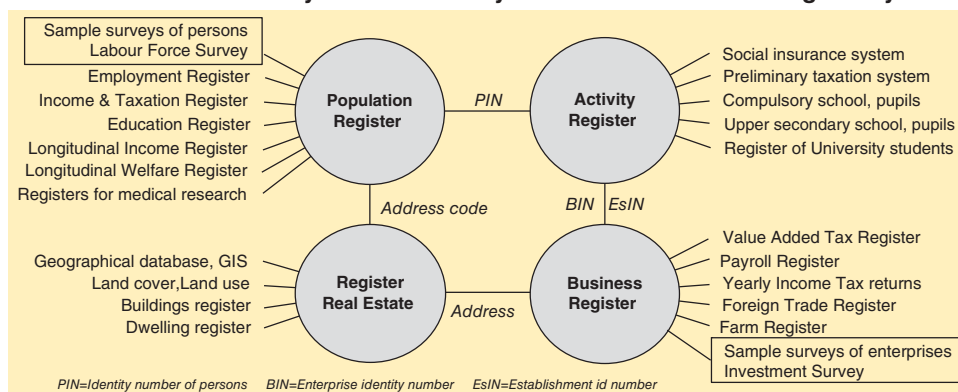
¹We will use the abbreviation NSO in all chapters.

The national statistical system can also be defined as the system of all censuses, sample surveys and register surveys that is carried out by the actors in Chart 1.1.

Chart 1.2 shows several register surveys together with two examples of sample surveys. The thin lines between the surveys indicate that microdata can be linked with identity numbers.

The four registers in the grey circles play an important role in this kind of system. They define populations of different statistical units and link these populations with each other. These four registers are called the *base registers* in the system.

Chart 1.2 The Statistical System in a country can consist of the following surveys:



Survey design refers to the development of methods that should be used for a specific survey. As a rule, the term is used for the design of sample surveys; but in subsequent chapters we discuss it with reference to the design of register surveys.

Survey system design is a term introduced by Laitila, Wallgren and Wallgren (2012) and describes the simultaneous work of improving or redesigning a system of surveys. For example, when a statistical population register has been developed in a country, all area samples of households can be replaced by samples of persons drawn from frame populations created with the new population register. This means that the whole system of household sample surveys is redesigned.

1.2 The traditional census-based system

Countries with a mainly traditional statistical system use interviewers to do population and housing censuses every ten years. Supported by maps and address lists, the interviewers go out and knock on all doors in the country. Sample surveys are conducted as a complement to the census, where interviewers go out and knock on an area sample of doors. The census is used to create the sampling frames.

Note that the *geographical location* of a person determines if that person is included in a sample or not. During the census or the sample survey interview, *all variables* required in the census or sample survey are collected by the interviewers. That is why concluding data sharing and cooperation agreements is not necessary – the different actors in the statistical system can manage their surveys independently. The interviewers can ask for names, birthdate and birthplace, but identities are not used in the production of statistics.

Census costs and value of information

A population and housing census is a costly and difficult operation, especially for developing countries. If the census has been successful, we obtain detailed information for small geographical areas and small categories of the population. However, the census estimates will be outdated after a few years. Using Swedish data, we can compare the information from a traditional census-based system with the information we obtain from a register-based statistical system.

Assume that we conducted a census in Sweden in 2001 and 2011 and that the next census will be done in 2021. Charts 1.3 and 1.4 show the number of employed persons in a municipality according to available data during February 2019.

Chart 1.3 Employment with census data

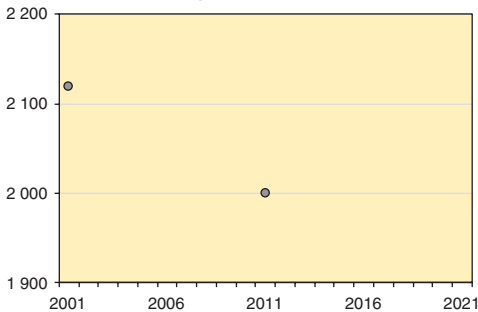


Chart 1.4 Employment with register data

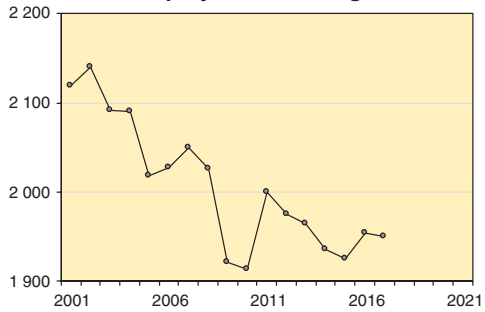


Chart 1.3 shows that there was a marked decline in employment between 2001 and 2011. But we do not know *when* the decline took place and we do not know what is happening *now*. We also know that the census involved substantial costs.

Chart 1.4 clearly shows the negative trend between 2001 and 2011. We can also see the effect of the financial crisis during 2009–2010. Finally, we can see what is happening now up to 2017. Chart 1.4 is based on estimates from Statistics Sweden's Employment Register and the costs are small compared with the costs for a census.

Remark: Administrative systems cover all time; censuses only give snapshots every tenth year.

In the early 1960s, the Nordic countries had statistical systems with area sampling and traditional population and housing censuses. Step by step, these countries managed to transition into completely register-based statistical systems; and in 2011 all countries conducted completely register-based censuses.

Chart 1.5 The year of establishing new statistical registers in the Nordic countries

Register	Denmark	Finland	Norway	Sweden
Population Register	1968	1969	1964	1967
Business Register	1975	1975	1965	1963
Dwellings Register	1977	1980	2001	2011
Education Register	1971	1970	1970	1985
Employment Register	1979	1987	1978	1985
Income Register	1970	1969	1967	1968
Population and housing census	1981	1990	2011	2011

Source: UN/ECE (2007)

We have visited many countries in Latin America and the Caribbean. All these countries had traditional census-based systems and the national statistical systems were decentralised. The national statistical office was responsible for the population and housing census, the agricultural census and some household sample surveys; the central bank was responsible for economic statistics and the National Accounts; and several ministries produced their own statistics.

All household sample surveys were based on area frames. Identities were not used when producing statistics and there was no tradition of data sharing and cooperation. If a country with this kind of statistical system wants to start using administrative registers for statistics production, then a comprehensive survey system redesign must be carried out.

1.3 New sources: Administrative registers and big data

Public authorities in all countries develop and maintain administrative systems that generate large volumes of microdata. When a national statistical office starts using administrative registers, a new production factor is added to the production system. The opportunities availed by the old methods of conducting a census or a sample survey remain, but new opportunities will be added to the production system when administrative registers are now used for statistical purposes.

The new administrative sources can often be difficult to use due to quality problems. However, it is still worthwhile to search for opportunities. In the future, new and better systems will be developed. National statistical offices should be active in the work to improve national administrative systems so that more and better systems can be used for official statistics.

Chart 1.5 clearly shows that the transition from the traditional census-based statistical system into the new register-based system took many years. All surveys changed when registers become available to produce official statistics.

- All censuses, sample surveys and register surveys will include identities. It will be possible to combine registers with other registers and to combine sample surveys with registers. Data can then be used much more efficiently, and this opens new possibilities for quality assurance and improvements.
- When different sources are compared, differences in coverage and differences between related variables are indicators of quality problems. The data sources need to be audited for quality.
- Censuses and sample surveys can use registers to generate frames. However, if the registers have coverage problems, it is wise to continue with area sampling. Area samples can then be used for quality assurance of registers.

Big data

Administrative registers are related to the more recent *big data* topic. Administrative data are considered by some authors as one kind of big data. However, we prefer to consider administrative data as a distinctive category, as they are much better structured and precisely defined in contrast to other types of data sources. In addition, administrative data has its own survey methodology, which is becoming established in more and more countries. If so-called ‘big data’ contain identities that can be linked to persons, areas, or enterprises, then big data are administrative data that can be linked with the system of statistical registers developed by the national statistical office.

Example: Toll payments and road sensor data

We sometimes drive over a new bridge and must pay a toll of 5 SEK (\approx 0.5 USD) every time we use the bridge. Cameras read the registration number of our car, and we subsequently receive a mail based on the car owner’s personal identity number with a request for a monthly payment for all the times we have used the bridge.

The request is based on a combination or system of three administrative registers: the register with camera data, the Vehicle Register and the Population Register. Thousands of registrations are made in this way every day and this is probably what many call ‘Big data’. But the 5 SEK payment is an example of a tax payment and should be linked with the Income Register at Statistics Sweden, where the important variables *disposable income of persons* and *disposable income of households* are created for official statistics.

There are many examples of toll systems and road sensor data that collect data regarding activities connected with vehicles. Since the registration numbers of the vehicles can be linked to the owners’ identities, all the data can be combined with other sources and used for statistical purposes.

All data in the register system in Chart 1.2 can be georeferenced and used to produce estimates for small geographical areas. These estimates can be supplemented with sensor and mobile phone data for the same areas.

1.4 Basic concepts and terms

The development of register-based statistics requires a common and rich register-statistical language. A common language within the theory of survey sampling is taken for granted. Terms such as frames, estimators and standard errors are well known and have a clearly defined meaning. Register-based statistics have the same need for well-established terms to stimulate the exchange of knowledge.

Two principles form the basis of this book – the *survey approach* to administrative data and the *system approach*. The survey approach involves the discussion of estimates, estimators and quality as in a book on sample surveys. The system approach builds on the *register system* concept. We also discuss the *production system* at a national statistical office and the role of administrative registers in the design and development of that system.

1.4.1 What is a register?

An administrative register is maintained to store observations on all objects to be administered; and the administrative process requires that all objects can be identified.

The following definition is valid for *administrative* registers:

An *administrative system* continuously generates *new data* to an administrative register; or it generates new administrative registers periodically.

An *administrative register* aims to include all the objects in a defined group of objects: the administrative object set. However, data on some objects can be missing due to quality deficiencies.

Data on the *object's identities* are used in the administration of objects. Therefore, the register can be updated and expanded with new variable values for each object.

Generation of new data, complete listing and known identities are therefore the characteristics of an administrative register.

Catalogue, directory, list, register, registry are different terms for the same concept. We use only the term *register*.

The following definition is valid for *statistical* registers:

A statistical register has been created by statisticians who use available administrative and statistical registers.

Complete listing and *identities* are also characteristics of statistical registers.

The administrative object set is replaced by the statistical concept *population*; and the known identities should be replaced by *anonymized identity numbers*.

The following are examples of registers:

- Civic, civil or national registration of the population in a country results in registers of citizens, births and deaths. This is an administrative system that continuously generates new data regarding the demographic events that affect the population.
- Income self-assessments from persons result in registers of all taxpayers for a given year. This is an example of an administrative system that generates new administrative registers yearly.
- In Sweden, enterprises with a turnover of SEK 40 million or more should report value-added tax monthly. This results in monthly VAT registers of reporting enterprises. For smaller enterprises, we obtain quarterly or yearly VAT registers. In all, we obtain three registers for three object sets: enterprises reporting monthly, quarterly and yearly.
- All export and import transactions are registered by Customs. Monthly registers are created with all transactions for a specific month. These transactions include identity numbers of exporting and importing enterprises.

The identities used in register processing can either be identity numbers that are unique within a national administrative system or an identity number in a subsystem with keys to the identities in other systems (as vehicles in the example with toll payments have links to the owners). It is also possible to use identities defined by, for instance, name, address, date of birth and place of birth.

1.4.2 Databases, records and observations

When Statistics Sweden migrated from mainframe computers to database servers, old terms such as *flat files* with *records* and *positions* were replaced by the term *database tables* with rows and columns.

We discuss these terms using the following example with data from an imaginary statistical register. Assume that we have a register containing data on all enterprises at a certain point in time. The number of objects in the register, illustrated in Chart 1.6, is given by N ; and the register contains six variables.

In a *data matrix*, statistical data are sorted so that the matrix columns are the *variables*, and the matrix rows are the *observations* for the objects. The register in Chart 1.6 is represented by a data matrix with N rows and six columns.

Every statistical survey (census, sample survey or register survey) aims to create one or several data matrices containing *microdata*, which will then be processed for statistical purposes. The term *data matrix* can be considered a statistical concept for such a data set.

The columns in the matrix contain measurements of variables; the rows in the matrix contain *observations* for the objects in the register. The six-dimensional observation for Object 2 has been marked in grey in the chart.