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

Edition History
John Wiley & Sons Ltd (2e, 2014)

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.

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

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.

Register-based Statistics: Registers and the National Statistical System, Third Edition. Anders Wallgren and Britt Wallgren.

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 Office[1] (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.

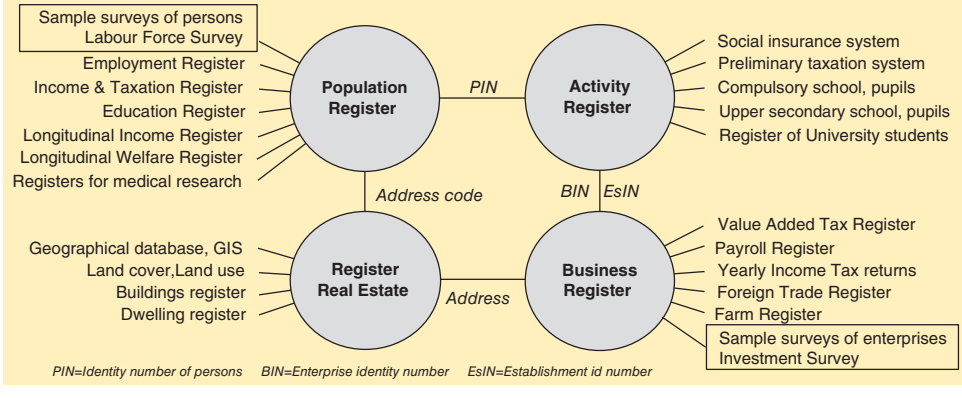
[1]We will use the abbreviation NSO in all chapters.

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

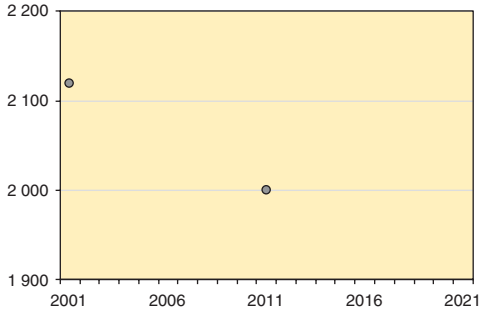


Chart 1.4 Employment with register data

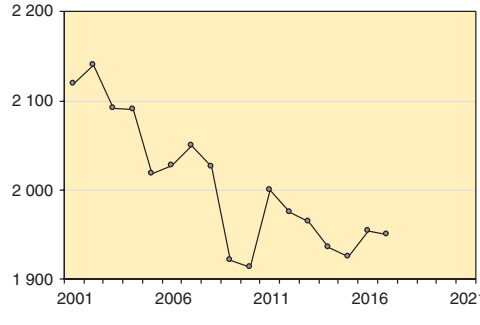


Chart 1.3 shows that there was a marked decline in employment between 2001 and 2011. But we do not know *when* the decline took place and we do not know what is happening *now*. We also know that the census involved substantial costs.

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 (≈ 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.