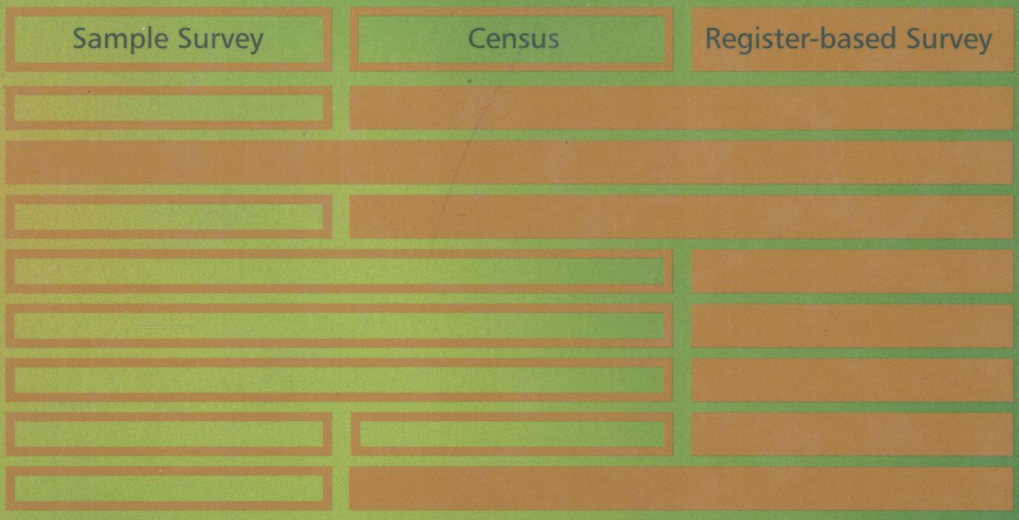
WILEY SERIES IN SURVEY METHODOLOGY

Register-based Statistics

Administrative Data for Statistical Purposes

Anders Wallgren
Britt Wallgren



WILEY

Register-based Statistics

Administrative Data for Statistical Purposes

Anders and Britt Wallgren
Statistics Sweden, Sweden

John Wiley & Sons, Ltd

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

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

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

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, ONT, L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

British Library Cataloguing in Publication Data

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

ISBN-13 978-0-470-02778-3 (HB)
ISBN-10 0-470-02778-9 (HB)

Typeset in 10/12pt TimesNewRoman by the authors.
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

Register-based Statistics

Preface

Register-based surveys are becoming more and more common within a growing number of national statistical offices, but are also common within enterprises and other organizations, where data from the organization's own administrative systems are used to produce statistics on, for example, production, sales and wages.

Although register-based statistics are the most common form of statistics, no well-established theory in the field has existed up to now. There have been no well-known terms or principles, which have made the development of both register-based statistics and register-statistical methodology all the more difficult. As a consequence of this, ad hoc methods have been used instead of methods based on a generally accepted theory.

Many countries are investigating the possibilities to use more and more administrative data for statistical purposes. It is necessary to reduce response burden and costs; increasing nonresponse in censuses and sample surveys also make this new strategy necessary. A new approach is necessary and register-based surveys require that suitable statistical methods be developed.

We have studied the requirements for register-based statistics through analysis of Statistics Sweden's system of statistical registers. During more that ten years, we have devoted an increasing part of our work, at the Department of Research and Development at Statistics Sweden, to the study of register-based surveys. We have also worked together with a number of manufacturing enterprises and analysed their administrative data for the purposes of their management. These experiences are also used in this book.

One purpose of the book is to describe the register system and discuss how it should work, presenting the possibilities offered by a functioning register system. Another purpose is to structure and describe the register-statistical methodological work which will provide the basis for the creation of the system. In several cases, we suggest new terminology. The book can be considered as a first step to a more systematic way of working with register-statistical issues. Our hope is that it will stimulate further development in this important area and encourage an overview of and improvements to the current ways of working. The necessary further development, among other things to ensure a better consistency between the different register-based surveys, is a task that may take many years.

Professor Carl-Erik Särndal has been a very important discussion partner during our work with the book. His broad experience from statistical offices in different countries and his background as a specialist in sample surveys have been enormously useful. In addition, around 50 persons within Statistics Sweden have read and commented on different parts of the first Swedish version of this book. Several individuals have also been interviewed to provide material for different examples and methodological sections.

Register-based Statistics – Administrative Data for Statistical Purposes A. Wallgren and B. Wallgren

The Swedish book was published 2004, and has been used in a number of study circles within Statistics Sweden. These study circles were very stimulation and have helped us in our work with this English version.

When preparing this version of the book we have used many valuable comments and suggestions from five anonymous reviewers engaged by John Wiley & Sons, Ltd.

Section 11.6, *IT Systems for Register-based Statistics*, is written by Lars-Göran Lundell, Statistics Sweden, and we have discussed many parts in the book with him to get the IT perspective on the register system.

The comments and encouragement from all these people mentioned above have been of great help in the work with the book.

It is our hope that *Register Statistics – Administrative Data for Statistical Purposes* and its proposals will stimulate the discussion on register statistics.

Örebro, Sweden
January 2007

Anders Wallgren
Britt Wallgren

Contents

Preface			**ix**
Chapter 1	**Register-based Surveys – An Introduction**		**1**
	1.1	Do we need a theory on register-based surveys?	1
	1.2	What is a statistical survey?	3
	1.3	What is a register?	4
	1.4	What is a register-based survey?	4
	1.5	Administrative and statistical information systems	10
	1.6	Why use administrative data for statistics?	12
	1.7	An overview of this book	15
Chapter 2	**How to Structure a Register System**		**19**
	2.1	A register model based on object types and relations	19
	2.2	The system of base registers	23
	2.3	The register system as a whole	29
	2.4	Building and using the system	30
	2.5	Standardised variables in the register system	33
	2.6	Statistical register systems outside Statistics Sweden	34
Chapter 3	**A Terminology for Register-based Surveys**		**41**
	3.1	Terminology – different language	41
	3.2	Register terms	42
	3.3	Terms for different kinds of variables	49
Chapter 4	**Sample Surveys and Registers**		**59**
	4.1	How can sample surveys benefit by the register system?	59
	4.2	Combining register-based surveys and sample surveys	61
	4.3	Comparing sample surveys and register-based surveys	63
Chapter 5	**How to create a Register – The Population**		**67**
	5.1	How should register-based surveys be structured?	67
	5.2	Determining the research objectives	70
	5.3	Making an inventory of different sources	72
	5.4	Defining a register's object set	72
	5.5	Defining and deriving objects	84
	5.6	How to produce regional register-based statistics	88

Chapter 6	**How to create a Register – The Variables**		**91**
	6.1	Deciding the register's variable content	91
	6.2	Forming derived variables using models	93
	6.3	Editing and correcting register variables	100
	6.4	Creating longitudinal registers	112
Chapter 7	**Estimation Methods**		**115**
	7.1	Estimation in sample surveys and register-based surveys	116
	7.2	Register-based surveys – Fundamental estimation methods	117
	7.3	Using weights in register-based surveys	119
	7.4	Estimation using weights – calendar year registers	121
	7.5	Calibration of weights in register-based surveys	123
Chapter 8	**Calibration and Imputation**		**127**
	8.1	The nonresponse problem	127
	8.2	Estimation methods to correct for overcoverage	138
	8.3	Methods to correct for level shifts in time series	140
Chapter 9	**Estimation with Combination Objects**		**147**
	9.1	Aggregation errors	147
	9.2	Estimation methods for multi-valued variables	149
	9.3	Linking of time series using combination objects	168
Chapter 10	**Quality of Register-based Statistics**		**173**
	10.1	Specific quality issues for register-based statistics?	175
	10.2	Errors in sample surveys and register-based surveys	176
	10.3	The users' and the producers' view of quality	181
	10.4	Detailed knowledge of a register's characteristics	182
	10.5	Overall appraisal of quality	190
	10.6	Main quality issues in different kinds of surveys	192
Chapter 11	**Metadata and IT-systems**		**193**
	11.1	Primary registers – the need for metadata	193
	11.2	Changes over time – the need for metadata	195
	11.3	Integrated registers – the need for metadata	196
	11.4	Classification and definitions database	197
	11.5	The need for metadata for registers	198
	11.6	IT systems for register-based statistics	200
Chapter 12	**Protection of Privacy and Confidentiality**		**209**
	12.1	Internal security	210
	12.2	Disclosure risks – tables	212
	12.3	Disclosure risks – micro data	216

Chapter 13 Coordination and Coherence **217**
13.1 Content-related coordination 217
13.2 Coherence 219
13.3 Consistent and coherent enterprise statistics 220

Chapter 14 Conclusions **227**

References **231**

Glossary **235**

Index **245**

CHAPTER 1

Register-based Surveys – An Introduction

This chapter and the next introduce a variety of concepts and principles that will be used in this book when discussing *register-based surveys,* that is, surveys that are based on data from administrative registers. These concepts and principles form the basis for a theory on this type of survey.

Register-based surveys are common within enterprises and other organisations, where data from the organisation's own administrative systems are used to produce statistics on for instance production and sales. Register-based surveys are also common at the national statistical offices in the Scandinavian countries, where many administrative registers are used to produce official statistics.

In this book, we will primary discuss register-based surveys at national statistical offices. There is an increasing interest in this area; many countries use more and more administrative data for statistical purposes and there is a growing demand for a theory on register-based surveys.

Our aim is to present statistical methods and principles of general interest, but we will use Scandinavian experiences and case studies from Statistics Sweden to illustrate these general methodological issues.

1.1 DO WE NEED A THEORY ON REGISTER-BASED SURVEYS?

Within the national statistical offices, three kinds of statistics are published – statistics based on sample surveys, statistics based on censuses and statistics based on administrative registers. It is most common to only differentiate between sample surveys and censuses, where the statistical office is responsible for the collection of the data. These two survey types are dominated by the work to collect data.

However, this book deals with the third type of statistics that are based on administrative registers where, instead of collecting data through surveys and censuses, administrative registers from different sources are adapted and processed to be suitable for statistical purposes. This kind of survey is called a *register-based survey.*

Sample surveys are based on methods that have been derived from an established theory – *sampling theory.* This theory has been developed both within the academic world and within statistical offices, and consists of terms and principles that are generally well known.

Register-based Statistics – Administrative Data for Statistical Purposes A. Wallgren and B. Wallgren

Scientific literature and journals develop and spread the methodologies for sampling and estimation. Because the terms and principles are well known, persons working with sample surveys can easily communicate and exchange their experiences.

Censuses with their own data collection are based on a long tradition of population censuses and the collection of data from local authorities, schools and different types of enterprises. Measurement errors, design of questionnaires and nonresponse are methodology issues that also apply for sample surveys. Censuses and sample surveys are closely related in terms of methodology – censuses are often considered as special cases where the sample is the entire population.

Statistics based on administrative registers will hereafter be called *register-based statistics*. Although this is the oldest and most common form of statistics, no well-established theory in the field exists. There are no well-known terms or principles, which makes the development of both register-based statistics and register-statistical methodology all the more difficult. As a consequence of this, *ad hoc methods are used instead of methods based on a generally accepted theory.*

One important reason for this shortfall is that the subject field of register-based surveys is not included in academic statistics. Statistical theory within statistical science is understood to consist of *probability theory* and *statistical inference.* Sampling theory is included within this theoretical school of thought, but register-based surveys based on total enumeration are not.

Unfortunately statistical science has so far not included any theory on statistical systems. Statistical offices, larger enterprises and organisations do not carry out separate surveys so often. It is more common that statistical information systems are built, which constantly generate new data. A statistical theory is necessary to describe the general principles and to develop the concept apparatus for such statistical systems. Register-based surveys should be included in this theory.

In 1995, Statistics Denmark published *"Statistics on Persons in Denmark – A Register-based Statistical System"*. The Danish book presents a systematic review of register-statistical work and describes how to design a well-prepared register system.

In this book, we build on and add to the Danish work. The next chapters introduce a number of register-statistical concepts and principles. The Glossary compiles all these concepts and terms. The aim is that all those working with the development of register-based surveys could then use the terms generally.

We formulate four principles for how administrative registers should be used:

Chart 1.1 Four principles on how to use administrative data

1. A statistical office should have access to administrative registers kept by public authorities. This right should be supported by law as the protection of privacy.
2. These administrative registers should be transformed into statistical registers. Many sources should be used and compared during this transformation.
3. All statistical registers should be included in a coordinated register system. This system will ensure that all data can be integrated and used effectively.
4. Consistency regarding populations and variables are necessary for the coherence of estimates from different register-based surveys.

We will use these principles in the book and gradually introduce the register-statistical terms that are needed for the discussions.

1.2 WHAT IS A STATISTICAL SURVEY?

The starting point for any survey is a number of questions in connection to a specific area of interest. A survey is carried out to try and answer these questions. The survey process can be described in more or less detail. Simply described, the work consists of the following phases:

1. Determining the research objectives and planning of the survey.
2. Procurement and processing of data.
3. Estimation, analysis of data and presentation of the results.

Within a national statistical office it is usual to work with surveys, which are repeated every year, quarter or month. With such surveys, work is mainly carried out in phases 2 and 3. However, these surveys have also had a phase of determining objectives and planning, even if this was a long time ago.

A separate survey can be a commission where the statistical office is to carry out the entire survey and this involves working with all the three phases. However, in many commissions, it is the customer who carries out phases 1 and 3 and the statistical office is only brought in to work with phase 2.

Phase 2 of a survey, the procurement of data, can be carried out in different ways:

a. With own data collection using a *sample survey.*
   Example: The Labour Force Survey is conducted in many countries. A new sample is taken monthly, with new data collection and reporting.
b. With own data collection using a *census.*
   Example: The traditional Population and Housing Census, in which all households and house owners are interviewed or asked to complete a questionnaire which is then processed by the national statistical office.
   Because censuses result in the creation of a register, microdata from censuses are also included in the system of statistical registers and can therefore form the basis for register-based surveys.
c. Existing microdata is used for a *register-based survey.*
   Microdata refer to data on individual *objects*. Existing administrative or statistical registers with data that, for example, refer to individual persons or enterprises are used for the purposes of the register-based survey.
   Example: In Section 1.4 below we give two examples of how statistical registers are created to meet the needs of different register-based surveys.

Because these three types of surveys differ in terms of methodology, it is appropriate to differentiate them conceptually. Sample surveys, censuses and register-based surveys are the most important types of surveys at a national statistical office.

A statistical population consists of N *objects* or *units* or *elements.* Of these three synonyms we will as a rule use the term *object* in this book.

1.3 WHAT IS A REGISTER?

An *administrative register* is maintained to store records on *all* objects to be administered and the administrative process requires that it is possible to *identify* all objects. The following definition is valid for both administrative and statistical registers:

> A *register* aims to be a complete list of the objects in a specific group of objects or population. However, data on some objects can be missing due to quality deficiencies. Data on an object's identity should be available so that the register can be updated and expanded with new variable values for each object. Complete listing and known identities are thus the important characteristics of a register.

The identities used in register processing can either be identity numbers who are unique within a national administrative system or an identity number in a subsystem with keys to the identities in other systems. It is also possible to use identities defined by for instance name, address, date of birth and birthplace.

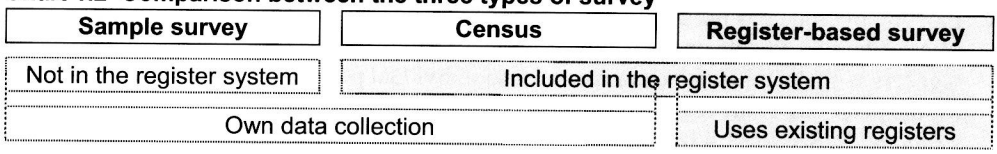
These identities will be used in exact matching of the objects in different registers, where the aim is to find identical or related objects in two registers.

A *statistical register* is based on data from administrative registers that have been processed to suit statistical purposes. The register processing which transforms administrative data into statistical registers gives rise to important methodological questions that will be discussed later in the book.

The term *statistical register* is used to describe registers within a system of statistical registers within a statistical office or other organisation. Such registers can be based either on a census carried out by the agency or on administrative registers from authorities and organisations outside the statistical office.

Data collection in a sample survey does not give rise to a register, as the micro data about the sample only consists of a small part of the surveyed population. Chart 1.2 compares the three types of survey that dominate at national statistical offices.

Chart 1.2 Comparison between the three types of survey



The term *register-based statistics* refers to statistics that are based on register-based surveys. When we discuss the register system, as in Chapter 2, we do not differentiate between censuses and register-based surveys. However, when we discuss methodology issues, the term only refers to register-based surveys.

1.4 WHAT IS A REGISTER-BASED SURVEY?

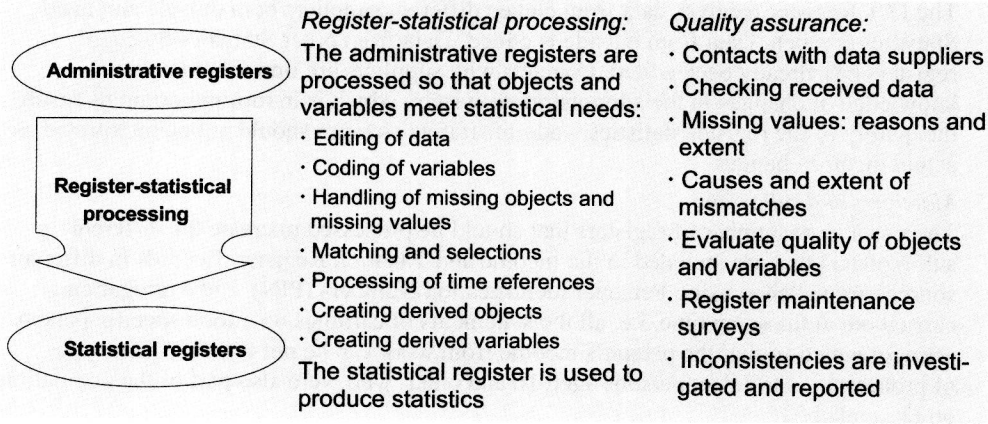
Administrative registers are created and delivered to a national statistical office

The original data formation is carried out in the authorities and organisations. The definitions of objects and variables are adapted to administrative purposes. Every authority

carries out controls, corrections and other processing that are suited to their administrative aims. When an authority delivers data to a national statistical office, further selections and processing may be carried out to meet the needs of the statistical office. The respective authorities also have metadata in the form of information on the definitions, data formation and quality. This type of information is also important for those receiving the data within the statistical office.

What happens when data is delivered to a statistical office as Statistics Sweden?
It is generally not a good idea to produce statistics directly from the received administrative registers because these are not adapted to statistical requirements. The object sets, object definitions and variables need to be edited and it will often be necessary to carry out some processing so that the register fulfils the statistical requirements for objects and variables. The register-statistical processing, which aims to transform one or several administrative registers into one statistical register, should be based on generally accepted *register-statistical methodology*. These methodological issues are discussed in more detail in the following chapters. The chart below shows the different elements included in statistical methodology work.

Chart 1.3 From an administrative register to a statistical register



In the next two subsections, we describe how two statistical registers are created. The examples are from Statistics Sweden, but they illustrate general principles. Each of these registers is created to meet the needs of a number of register-based surveys. The examples illustrate how administrative data are transformed to meet statistical needs and how the system of statistical registers (at Statistics Sweden) is used when creating statistical registers. The main part of the work with a register-based survey is the work spent on creating an appropriate register.

1.4.1 Statistics Sweden's Income and Taxation Register

This register utilises many administrative sources. Many administrative variables are used to create important statistical variables. Besides these administrative sources it is necessary to use the register system at Statistics Sweden: the Population Register is used to define the population of the Income and Taxation Register, and important classification variables are imported from other registers in the system to the Income and Taxation Register.

1. *Data formation at the National Tax Board*
   The annual income assessment is based on tax declarations from income earners

and the taxation decisions of the local tax authority. Both the income earner and the tax authority use statements of earnings regarding salary, sickness benefit and interest that the employers, social insurance office and finance companies are responsible for. The National Tax Board ultimately compiles this information. Declarations, statements of earnings and taxation decisions can be changed and supplemented. Data for one person can thus be very complex.

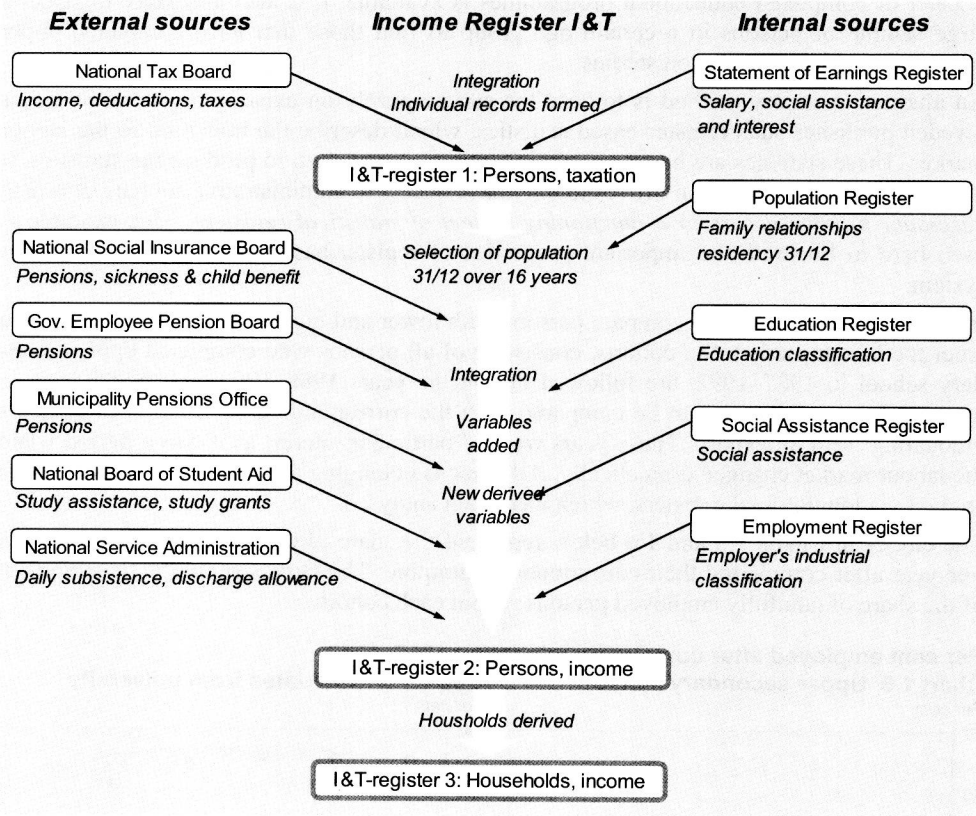
2. *Microdata deliveries to the Income and Taxation Register*
   The Swedish National Tax Board annually creates databases that contain information on Sweden's population. The data files for one year – containing around nine million records, each with around 300 variables – are delivered directly to the Income and Taxation Register at Statistics Sweden.
3. *Metadata to the Income and Taxation Register (I&T)*
   Record descriptions with variable names and variable definitions accompany the deliveries from the National Tax Board. Tax declaration forms, statement of earnings forms, taxation decisions, tax declaration instructions and instructions to employers are also needed to be able to interpret the data.
4. *Editing of data*
   The I&T Register receives data from eleven different suppliers both outside and inside Statistics Sweden. Data from outside is edited. Data from other Statistics Sweden registers has already been edited. Contacts with suppliers are important to obtain knowledge of changes in the administrative system, which is in turn important to ensure the quality of the register statistics – administrative changes should not be interpreted as actual income changes.
5. *Matching and selections*
   There is a large number of registers that should be processed to create the different sub-registers that are included in the Income and Taxation Register. Records in different sources are matched using Personal Identification Numbers (PIN), and aggregation is carried out at the same time, i.e. all the statements of earnings data for a specific person are aggregated so that the person's income from work can be put together. One type of processing is to select persons aged 16 and older, who were also part of the population on December 31.
6. *Derived objects are created*
   More information on certain relations helps to form household units. Between adults, the relations *married* or *cohabiting adults with children in common* result in that they are placed in the same household unit. These relations are shown by the family members' personal identification numbers, these reference variables are found in the taxation data and in Statistics Sweden's Population Register.
7. *Derived variables are created*
   A large number of derived income variables are formed. For instance, the wage or salary amounts are aggregated from the different earnings data to become an individual's *income from work*. Every person's total income from work and capital plus transfer payments minus tax becomes the person's *disposable income*. For households, variables such as *household type, number of consumption units* and *disposable income* are formed.

The chart below shows how the Income and Taxation Register receives administrative data from a variety of different external sources and some Statistics Sweden registers. The term *source register* refers to the administrative sources and the Statistics Sweden registers that are used to create the new register. The different phases when the source registers are used

during the process to create the new statistical register are shown in the middle column in the chart.

This example shows the importance of the four principles in Chart 1.1. Statistics Sweden has access to many administrative registers with variables describing different kinds of income. The object set and the administrative variables have been processed to meet statistical needs. Many sources have been used to produce a statistical income register with rich content. The population in the income register is consistent with other statistical registers within the register system.

Chart 1.4 Different data sources for the Income and Taxation Register (I&T)



The Income and Taxation Register is an important part of Statistics Sweden's register system. It is used to describe the income distribution, for regional income statistics and it is also the basis for longitudinal income registers used by university researchers.

1.4.2 Longitudinal register – education and labour market

The Income and Taxation Register mentioned above is directly based on large amounts of administrative data. However, many important registers at Statistics Sweden are not directly based on administrative data; they are instead based on already existing statistical registers in the register system. The example illustrates how existing data can be used in a new and more advanced way after specially adapted register processing.

The entry of young persons into the labour market after completing their studies is nowadays an important area for different surveys. Such surveys should be carried out as *longitudinal surveys*, where groups of persons are followed over a period of years. If these surveys are carried out as sample surveys, a sample is taken every year among persons completing a specific educational programme and each sample is interviewed or asked to fill in a questionnaire once a year for a period of years, in this case seven years.

This survey method has its disadvantages, partly that the burden on the respondents is heavy – the selected individuals must answer a large number of questions every year – and partly that nonresponse will gradually increase over the period. In addition, if no adequate register of completed educational programmes is available, it is also necessary to select a large sample of persons in a certain age group to find those that have completed upper secondary or higher education studies.

An alternative survey method is to base the survey purely on existing registers. Statistics Sweden publishes such register-based statistics, which describe the transition to the labour market. These statistics are based on administrative sources but, to produce the statistics, it is not sufficient for statistical offices to only have access to administrative sources. *It is also necessary to have access to a functioning system of statistical registers.* This example is used here to illustrate the important properties of register-based statistics and a register system.

In the charts below, we can compare persons with lower and higher education as they try to enter the labour market. Six cohorts, consisting of all persons who completed upper secondary school in 1987–1992, are followed during the years 1988–1993 and their transition into gainful employment can be compared with the corresponding six cohorts of students graduating from university. These years were of particular interest as it was a period when the labour market changed dramatically. All persons belonging to these twelve cohorts were studied via longitudinal registers, which were then analysed.

The circles in Charts 1.5 and 1.6 below represent the share of gainfully employed persons *one* year after completing their educational programme. The curves show the development of the share of gainfully employed persons within each cohort.

Per cent employed after completing education 1987–1992

Chart 1.5 Upper secondary school

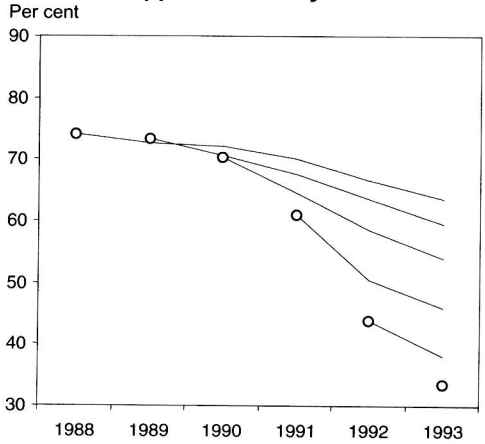
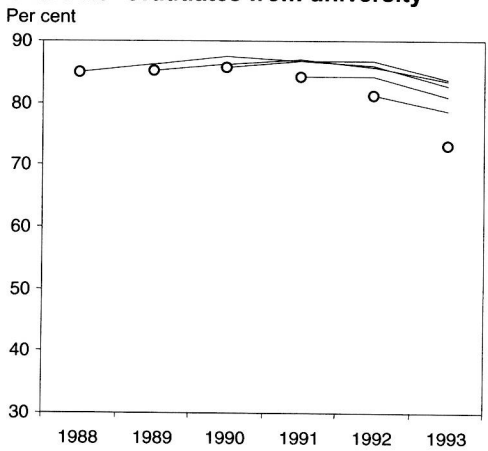


Chart 1.6 Graduates from university



At the beginning of the 1990s, the most serious crisis in the Swedish labour market since the 1930s occurred. Charts 1.5 and 1.6 show how the economic downturn at the beginning