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.

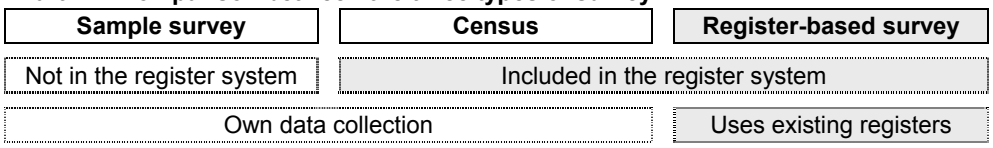
These identities will be used in exact matching of the objects in different registers, where the aim is to find identical or related objects in two registers.

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?

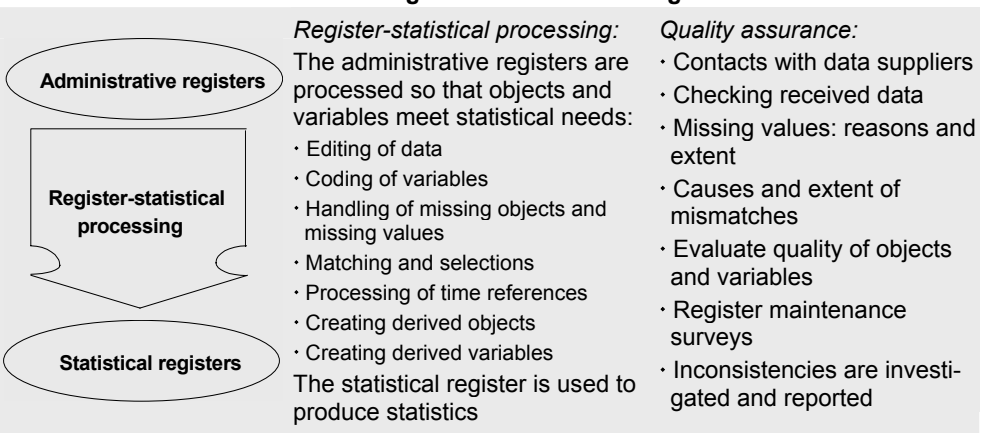
Administrative registers are created and delivered to a national statistical office

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.

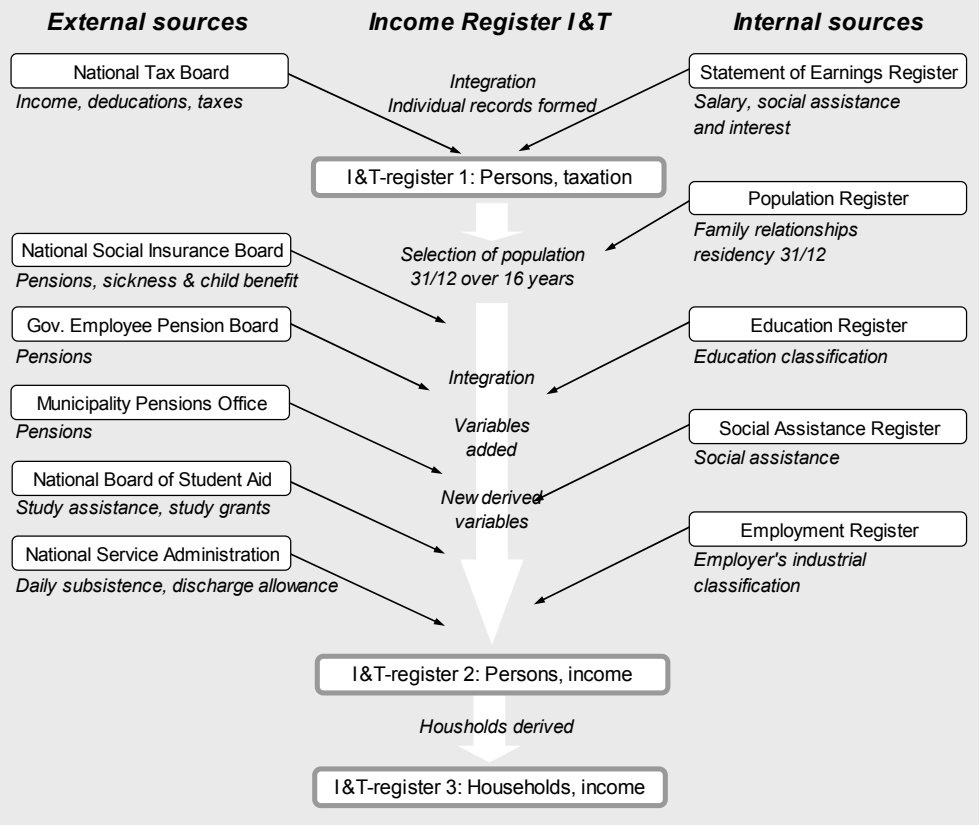
2. *Microdata deliveries to the Income and Taxation Register*
   The Swedish National Tax Board annually creates databases that contain information on Sweden's population. The data files for one year – containing around nine million records, each with around 300 variables – are delivered directly to 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

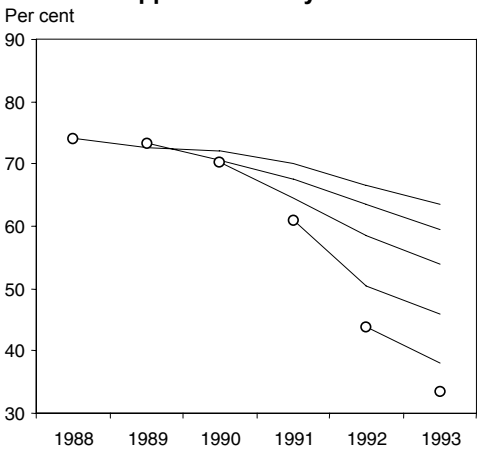
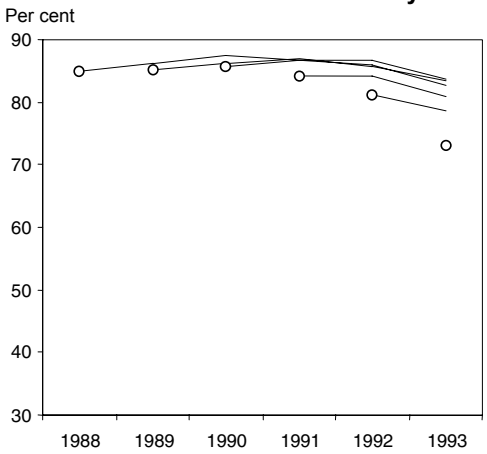


Chart 1.6 Graduates from university



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

of the 1990s changed the possibilities for young persons to enter the labour market. The share with gainful employment one year after completing upper secondary studies changed dramatically during these years. For those with university degrees, the development was not so depressing – this appears to show that higher education studies were worthwhile.

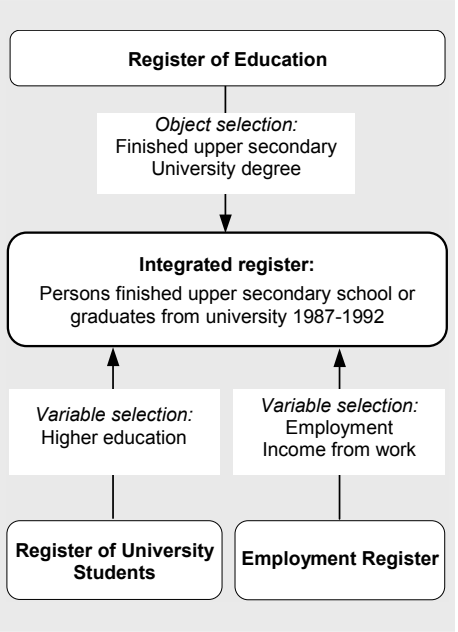
This example is based on longitudinal data. Every year, each person is classified as gainfully employed, studying in higher education or neither gainfully employed nor studying. For gainfully employed persons, the annual income from work is registered as well as information on which sector the person is working in. Persons are also classified by course/study programme, sex and region, which permit detailed reporting. The annual cost for this statistical product was SEK 0.4 million (approximately $ 50 000), half of which was printing cost.

How was this register-based survey carried out?

The charts above show the transition of young persons from education to the labour market. The longitudinal register that the charts are based on was created in the following way:

– By combining information from three Statistics Sweden registers for 1987–1993 (i.e. a total of 21 different registers), a new *integrated register* was created, which is marked in a bold frame in Chart 1.7 below.
– The objects in the new register were created by selecting certain objects in the Register of Education.
– Variable values are imported into the new register by *matching* the objects in the new register with the corresponding objects in the Education Register, the Register of University Students and the Employment Register. This is illustrated in the chart below.

Chart 1.7 Creating an integrated register for the transition to the labour market



Register of Education
Based on data from schools or universities
Object: Person
Variables: Personal identification number, education, year of degree
Processing: Selection of persons who finished upper secondary school or graduated from University

Register of University Students
Based on data from universities
Object: Person actively studying
Variables: Personal identification number, pursuing higher education
Processing: Matching against selected group. Variables are selected and transferred to the integrated register

Employment Register
Based on statements of earnings, income, population and business registers
Object: Person
Variables: Personal identification number, employment status, income from work
Processing: Matching against selected group. Variables are selected and transferred to the integrated register

We have mentioned here a variety of important terms. By bringing together or integrating information from several registers, an *integrated register* is created. We distinguish between *object selection* where certain objects are selected and *variable selection* where

certain variables are selected. *Matching* means that the links in two or more registers are compared, resulting in a match or mismatch.

To create the new integrated register, no new administrative sources were used. The register was instead based on the increased usage of existing registers. The Employment Register is also based on existing registers within Statistics Sweden's register system. The Register of University Students, however, is built from administrative data that Statistics Sweden receives from authorities outside Statistics Sweden.

To create and coordinate these registers, extensive processing of register data in several registers is necessary. This means that having access to administrative sources is not sufficient to be able to produce these statistics. Without Statistics Sweden's coordinated system of statistical registers, it would be impossible to produce these statistics. When the register system is used, exact matching of records in different registers is done with standardised identification numbers. In the example above, Personal Identification Numbers (PIN) are used.

1.5 ADMINISTRATIVE AND STATISTICAL INFORMATION SYSTEMS

Using administrative data for statistical purposes is not something specific to Statistics Sweden or to statistical offices in other countries. It is also a common practice in large enterprises and organisations. Administrative systems are used generally as sources for statistical information and there is no major difference between the following enterprise example and register-based statistics at a national statistical office:

– Statistics on staff and salaries within an enterprise can be produced using the personnel management system.
– Population and income statistics are produced at a statistical office using data from the National Tax Board's tax collection system for population registration and tax assessment.

Register-based surveys have become more and more common within enterprises and organisations. Knowledge about register systems, register-based statistics and register quality is not only needed within a national statistical office but also in a more general sense. The following extract from a job advertisement shows this:

Market analyst

As an analyst in the marketing department, you will be an important cog in the wheel of our enterprise's continued growth. You will manage and develop the use of one of the enterprise's most valuable assets – our client register.

You will work with campaign analyses, drafting reports, segmenting and ensuring the quality of the register. You will maintain contact with external register systems and work closely with the marketing manager.

Certain information systems are built up solely for statistical purposes, such as the Labour Force Surveys, which are conducted in many countries. Such systems can therefore be completely designed according to statistical principles.

Other information systems are used both for administrative and statistical purposes, which can sometimes lead to conflicts with regard to the structure of the system. In general, these systems are primarily intended for administrative purposes and the statistical information is a by-product.

However, there are several differences between a pure administrative system and a pure statistical system. These two kinds of systems are compared below.

1.5.1 Different purposes

Information in an administrative system is used as a basis when taking administrative measures and decisions that will affect the objects in the system.

Example: A personnel management system is used to carry out salary payments.

Information in a statistical system is used as the basis for analysis and from the analysis one will draw conclusions. These conclusions can then become the foundation for policy-related decisions.

Example: A statistical salary system is used to study salary structure. How has this changed? What are the differences in monthly salaries between different categories of staff? This analysis could then involve a change in policy relating to salary issues, e.g. females should be paid better.

1.5.2 Different roles for individual objects

In an administrative system, decisions are made and measures are taken with regard to individual objects. To this end, information relating to that specific object is retrieved.

Example: Salaries are paid to every employee in an enterprise. Administrative information is checked and the salaries and taxes for the employees can be calculated.

In a statistical system, the individual objects are not interesting in themselves. In a statistical analysis aggregated estimates are calculated and compared for groups of objects.

Example: Salary totals, average salaries, the dispersion of salaries, etc. are calculated for the different categories of staff.

1.5.3 Approaches regarding errors

From an administrative point-of-view, certain items of the individual information must be absolutely correct, but other items can be more approximate. From a statistical point-of-view, errors can exist but they should be carefully controlled and attempts are made to reduce the errors, which can be significant for the statistical conclusions. Errors can be accepted in some data, but only if these are considered to have a limited effect.

Example: The personal identification number in a personnel management system must be completely correct from the point-of-view of salaries and tax administration. As the Swedish personal identification number contains the date of birth, it can also be used to describe the age structure of the staff. If, for example, 30% of the staff has an incorrect number for the month in their registration number, this would not affect the statistical analysis particularly although the salary and tax routines would become impossible.

1.5.4 How should administrative data be processed for statistical purposes?

A general principle is to combine many sources when a statistical register is created. There are many reasons for this: Variables from different sources can be used to achieve a rich content, as in Sections 1.4.1 and 1.4.2; other reasons are that coverage and editing possibilities can be improved.

When the object set in an administrative source is not suitable as a relevant statistical population, a number of sources should be combined to create an object set with good coverage.

Example: A business register at a national statistical office is based on administrative sources. With five sources we created a Business Register for Sweden containing all enterprises (legal units) active during 2002. Each source consists of the legal units in one taxation system. In the table below, undercoverage and overcoverage of the sources are compared with our final Business Register. Source 1 has the earliest and source 5 has the latest available information. The administrative object sets in each source is adequate for each of the five taxation systems. But taken alone, each source is of low *statistical* quality, however, if all sources are combined, the coverage is good.

Chart 1.8 Overcoverage and undercoverage in five administrative sources

	Source 1	Source 2	Source 3	Source 4	Source 5
Overcoverage	41%	0%	0%	0%	0%
Undercoverage	21%	74%	74%	30%	9%

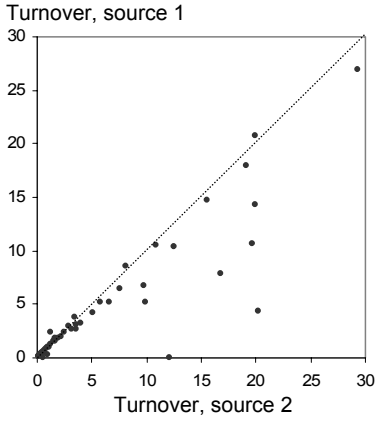
In sample surveys and censuses, editing as a rule uses the collected data only. In register-based surveys, however, it is possible to compare variables from different sources, and this gives better possibilities to find and correct errors. Measurement errors and other kinds of inconsistencies can be detected if different sources are compared.

Example: In Chart 1.9 the enterprise units (each consists of a number of legal units) in a Swedish business survey are compared with respect to turnover reported in two administrative sources.

Due to different definitions, source 2 should report larger turnover than source 1. This means that the four enterprise units above the diagonal line should be checked – probably these enterprise units are incomplete.

Also the enterprise units with large deviations under the diagonal line must be checked – probably the reporting in source 1 has been delayed for some legal units belonging to these enterprise units.

Chart 1.9 Comparing sources



1.6 WHY USE ADMINISTRATIVE DATA FOR STATISTICS?

There are both prejudices and legitimate criticism against register-based statistics. The prejudices consider statistics based on administrative data to be cheap but of bad quality, compared to 'true' survey statistics. The legitimate criticism can refer to relevance errors, comparability problems and that the statistical quality is not under control. Unfortunately, similar criticism can be directed towards sample surveys for which frame errors, measurement errors and nonresponse errors can be significant and undetected, irrespective of whether the sampling errors are under control.

Our answer to these types of comments is that statistics on society should consist of both register-based statistics and statistics based on data collected by a national statistical office. It is therefore not a question of which method is better than the other, but more that in certain situations register-based surveys are more effective and in others, sample surveys are most appropriate.

1.6.1 Pros and cons of register-based surveys

As we have mentioned, there is a common but often diffuse idea that statistics based on administrative data is of low quality. Is this idea justified for the administrative sources that a statistical office uses? A very large part of Statistics Sweden's register system is based on data from the administrative population register and tax administration. Would these statistics be of higher quality if Statistics Sweden collected the data itself in parallel with the National Tax Board's collection of administrative data? This is hardly the case – Statistics Sweden's own attempt to collect these data would be expensive, would increase the burden on the respondents and would likely produce data with more measurement errors.

Chart 1.10 Pros and cons of surveys based on data collection or registers

	Advantages	**Disadvantages**
Surveys based on data collection: sample surveys and censuses	Can choose which questions to ask Can be up-to-date	Some respondents .. ... do not understand the question ... have forgotten how it was ... do not respond (nonresponse) ... respond carelessly Burden on respondents can be high Expensive Low quality for estimates for small study domains (for sample surveys)
Register-based surveys	No further burden on the respondent for the statistics Low costs Almost complete coverage of population Complete coverage of time Respondents answer carefully to important administrative questions Good possibilities for reporting for small areas, regional statistics and longitudinal studies	Cannot ask questions Dependent on the administrative system's population, object and variable definitions The reporting of administrative data can be slow; the time between the reference period and when data are available for statistical purposes can be long Changes in the administrative systems make comparisons difficult Variables that are less important for administrative work can be of lower quality

In this table, we have summarised the advantages and disadvantages of the two survey methods. The significance of the disadvantages can vary in different survey situations.

Our example of the transition of young persons from education to employment can illustrate both the advantages and disadvantages of register-based statistics. We have already mentioned that register-based surveys are appropriate for longitudinal studies. Another advantage is that it is possible to report results for many sub-groups, different courses, study programmes and regions. But a disadvantage with certain types of administrative sources is that administrative systems often require a long time from the reference period to

when the data are available, which can cause delays with register-based statistics. In our example, the report describing the period 1987–1993 was published in June 1995, i.e. one and a half years after the end of the period.

So it is not just a question of *if* administrative data is to be used but also *how* it is to be used. Our response to *how* is that administrative data should in general not be used as it is, but should be processed for statistical purposes. An attempt should be made to build a system of coordinated registers – this can be advantageous both in terms of quality and cost, and quality assurance should be an important component of the system.

There are many development trends that threaten the quality of statistics based on sample surveys or censuses. The increased usage of telemarketing and number presentation on telephones means it is harder to carry out telephone interviews. If respondent motivation decreases, nonresponse and problems with measurement errors will increase. It will become harder to motivate the double provision of data – why respond to a questionnaire on the enterprise's turnover when you also submit a Value-Added Tax (VAT) declaration to the Tax Board which includes the same information? All these circumstances point to an increase in the significance of register-based statistics. Evidence that double provision of data to Statistics Sweden and to another authority is regarded as unreasonable can be seen in this local newspaper clipping:

> Mariestads Tidning, 26 June 2000 (*translated from newspaper article*):
>
> **Refuse to send statistics to Statistics Sweden!**
>
> Mr R from the B-farm thinks that the authorities should be able to find the information from their own registers.
>
> Mr R refuses to send in statistics to Statistics Sweden. Because he already sends in information every other week to the Swedish Board of Agriculture, he thinks that the authorities should cooperate with each other instead. …

1.6.2 The cost aspect – are register-based statistics cheap?

"It is quite clear that the Member States find themselves in the paradoxical situation of having to face a number of budget cutbacks at the same time as providing users with an increasing volume of high-quality relevant information." So began Yves Franchet, then Director General for Eurostat, a seminar (Eurostat, 1997) on the use of administrative sources for statistical purposes. The quote illustrates the need for more effective statistical systems. These requirements for increased efficiency can be met by combining two strategies:

- By using more administrative sources, the cost of data collection is reduced both for the respondents and the statistics producers.
- A more effective and flexible usage of existing data would mean that new requests could be met without the costly collection of new data. This could be achieved by using the administrative sources to create a system of coordinated statistical registers.

The construction of such a register system would be costly but, with increased use of the system, the marginal costs would decrease. In the example above of the transition of young persons from education to working life, the marginal cost of the product was small.

The burden on respondents is also a cost that can be reduced by moving from statistics based on data collection to register-based statistics.

Sample surveys are an increasingly expensive method; the number of telephone calls to first contact increase and the costs of reminding an increasing percentage of respondents who don't send back questionnaires are also rising.

The costs for sample surveys per inhabitant are higher for smaller countries – for corresponding domains of study; the same level of accuracy requires almost the same sample size in a small country as in a larger country. Therefore, it becomes especially important for smaller countries to build up a statistical register system.

1.7 AN OVERVIEW OF THIS BOOK

Every chapter in this book contains proposals for change. A new approach, new terminology and new methodologies are needed so that the register system and register-based statistics can be developed and function in an even better way than they do today. Below we provide a summary and overview of the contents of the chapters in the book.

Chapter 1 Register-based surveys – an introduction
So that register statistics can be developed, a *register statistical theory* is needed, with statistical systems at the centre. A well functioning register system forms the basis for the effective production of statistics.

Chapter 2 How to structure a register system
The register system's structure is presented here. The *register model* is an important tool to help spread understanding of the register system. The significance of *administrative sources* is discussed and the role of the four *base registers* is defined.

Chapter 3 A terminology for register-based surveys
A number of *register-statistical terms* are introduced; without good terminology, register theory would be vague and the exchange of experiences difficult. Terms for different kinds of registers are presented and variables are defined by their role in the register system. Variables derived via matching are also described.

Chapter 4 Sample surveys and registers
How can sample surveys benefit from the register system and how can sample surveys and registers be combined? An outline of these topics is given and also the differences between the methods used in sample surveys and register-based surveys are discussed. It is important to distinguish between surveys with different methodologies – on the one hand, sample surveys and censuses and, on the other hand, register-based surveys.

The data collection phase in surveys with their own data collection corresponds to the different types of register processing carried out to create a statistical register. This register processing should be studied from a statistical science point-of-view.

Chapter 5 How to create a register – the population
The procedure of creating a register is structured into five phases. For register-based statistics the term *register population* plays in important role, corresponding to the term *frame population* in sample surveys. All available sources should be used to create a register population with good coverage. By combining different administrative sources it is also possible to define a statistically relevant register population. A number of principles are presented, such as *'everyone should use the standardised populations from the base registers'* and *'everyone should support the base registers so that these have as high a level of quality as possible'*.

Chapter 6 How to create a register – the variables

Derived variables are central for register-based statistics. Administrative variables are used to define statistically important derived variables. Editing of administrative data is discussed via a number of case studies. There are some important differences between editing of data in a sample survey and editing of administrative data – consistency editing of many sources at the same time, and searching not only for variable errors but also for object errors.

Chapter 7 Estimation methods

Estimation methods are also necessary within register-based statistics. Simply summing the data is not always sufficient. The traditional approach within register statistics is that no estimation methods are needed, but the same statistical approach that are currently used by those working with estimation problems for sample surveys and censuses should also be applied to register statistics.

We differentiate between basic and supplementary estimation methods. The supplementary methods use weights in a similar way as weights are used in sample surveys. How weights are used and the calibration of weights is illustrated by examples.

Chapter 8 Calibration and imputation

Different estimation methods can be developed to deal with problems due to nonresponse or *missing values, overcoverage,* and *level shifts in* time series.

When different registers are integrated and variables are imported from one register to other registers, quality flaws such as missing values are also imported to these other registers. This means that it is not sufficient to adjust for missing values in a variable only in one register, the adjustment method must adjust for missing values in this variable in the whole register system in a consistent manner.

Overcoverage can cause serious errors in register-based statistics. We propose that calibration of weights can be used to adjust for overcoverage in the base register where the problem exists. All other statistical products using the base register will then use these weights.

Calibration of weights can also be used to adjust for level shifts in time series at the micro level. With these weights, consistent and linked time series can be produced.

Chapter 9 Estimation with combination objects

Multi-valued variables, such as Industry, are today used in a way that gives rise to aggregation errors. Special estimation methods using weights are introduced in this chapter that can be used for multi-valued variables to reduce these aggregation errors.

Combination objects are introduced for the estimation issues that are related to multi-valued variables. Such combination objects can also be used to adjust for level shifts in time series at the micro level.

Chapter 10 Quality in register-based statistics

Sampling errors have for a long time been regarded as the most important error in sample surveys. In register-based surveys there is no sampling phase; instead this kind of survey is dominated by the *integration phase*, where data from different sources are integrated into a new statistical register.

During the integration phase the register population and derived objects are created, variables are imported from different sources and derived variables are created. The kinds of errors that have their origin from the integration phase should be called *integration errors*.

Specific *quality indicators* are needed that suit the needs of register-based statistics. A number of indicators are presented and, for a specific register, the important indicators for that particular register are selected. Additionally, an *overall assessment* of quality should be made.

Chapter 11 Metadata and IT-systems

Register statistics requires a metadata system in which large amounts of *formalised metadata* can be processed using the appropriate IT tools. A *calendar* with the important changes and a *definitions database* are important parts of such a system.

The IT system for register statistics should be documented in a different way to the traditional *systems documentation*, which relates to surveys with their own data collection. *Data Warehouse technology* can be a tool for:

- more effective register management,
- an increased amount and more uniform metadata,
- simpler and more secure retrievals, and
- a better overview of the system's content.

Chapter 12 Protection of privacy and confidentiality

Well functioning routines for *the protection of privacy* are a very important part of the register system. Minimising the use of variables with information in plain language and official identification numbers should be considered. Routines for the *protection of disclosure* should always be included in the work to present and disseminate tables based on register statistics and when micro data are released for research purposes.

Chapter 13 Coordination and coherence

The concept *coherence* refers to that estimates from different surveys can be used together. For example, for a ratio to be meaningful, the numerator and the denominator must be coherent. Consistent surveys give coherent estimates.

Statistics from different sources can be consistent, i.e. have a high level of coherence through:

- ensuring consistency regarding *populations* (relating to definitions of both object and object set),
- ensuring consistency regarding *variables,* and
- using calibration methods that give consistent *estimates*.

An example with business statistics is used to illustrate a method for coordinated and consistent statistics. The aim is to show how inconsistencies can occur between surveys carried out at different points in time.

Chapter 14 Conclusions

In the last chapter we draw some general conclusions from the previous chapters. A new approach towards administrative data is necessary and development of register-based statistics should be recognised as an important field for statistical science.

References

Glossary

Index