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.

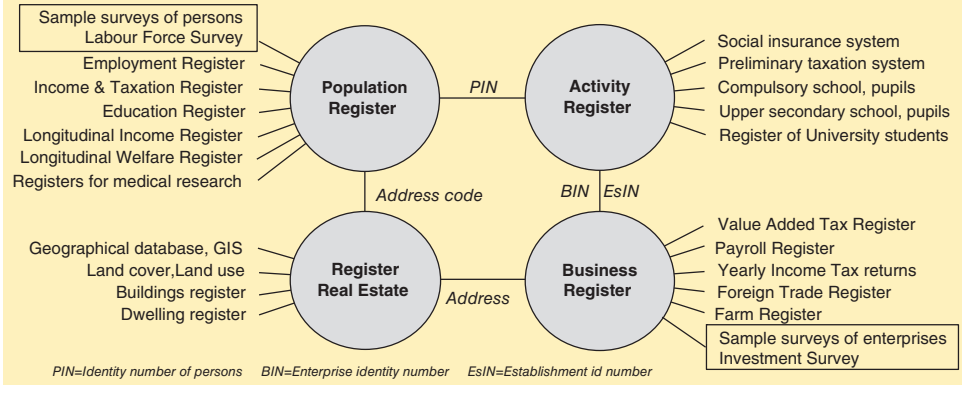
[1]We will use the abbreviation NSO in all chapters.

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

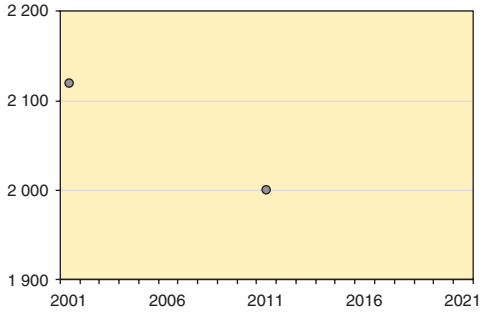


Chart 1.4 Employment with register data

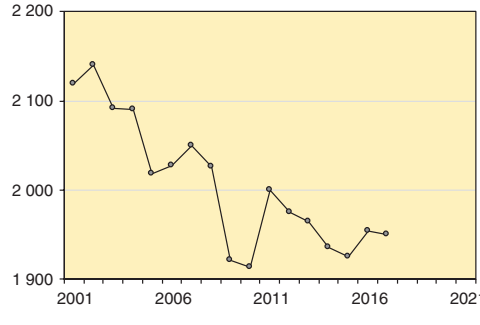


Chart 1.3 shows that there was a marked decline in employment between 2001 and 2011. But we do not know *when* the decline took place and we do not know what is happening *now*. We also know that the census involved substantial costs.

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.

Chart 1.6 Example of a register and data matrix

	Variable 1 **Name of enterprise**	*Variable 2* **Address**	*Variable 3* **Business identity number**	*Variable 4* **Turnover**	*Variable 5* **Employees**	*Variable 6* **Industry**
Object 1	A's Painters	Address 1	BIN 1	12	9	F
Object 2	B's Bakery	Address 2	BIN 2	3	4	D
-	-	-	-	-	-	-
Object N	Z's Factory	Address N	BIN N	211	76	D

The old term *record* generates misunderstandings. It is a synonym to the term *observation* or *row* in a data matrix. But we often see that people use the term *administrative records*, and we suspect that they mean *administrative registers*. This becomes even more complicated in Spanish. The Spanish word *registro* stands for both *register* and *record*. We suggest that the term *record* should be replaced with the term *observation*. And if you mean *administrative registers*, state that and do not state *administrative records*.

The terms database and database table

Data matrices such as those shown in Chart 1.6 are stored in databases, which use various IT terms that are described here. In this book, the discussion is always on a conceptual level, namely how it is *thought* that registers and data matrices will *look*. Conceptual models often differ from the physical implementation of a database. Chart 1.6, for example, shows complete agreement between the register, the data matrix and the database table. However, this may not be the case in an actual systems solution.

The aim of a systems solution is that the database should be a flexible base for many uses in the register system. The structure of the data matrix can also be saved in a so-called *view* that shows you the data matrix that can differ from the way that the data has been stored.

Charts 1.7 and 1.8 illustrate different ways of describing the same database. Chart 1.7 shows the traditional way of describing a database. Chart 1.8 shows an example of a database's contents. The charts show the structure of the systems solution, i.e. how the data are physically stored. The database is *normalised* to ensure optimum consistency and space-effective storage.

Chart 1.7 A conceptual database model with three database tables

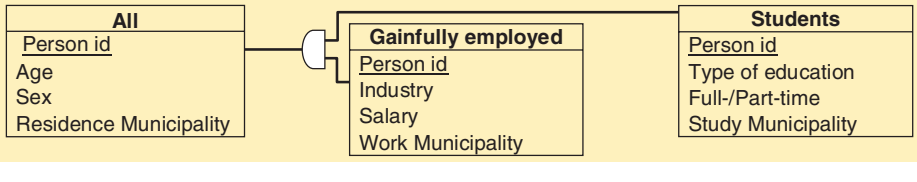


Chart 1.8 A database on individuals with three database tables

All

Person id	Age	Sex	ResMun.
PIN1	20	F	0586
PIN2	23	M	0586
PIN3	31	M	0586
PIN4	32	F	0586
PIN5	33	M	0586
PIN6	40	F	0586
PIN7	59	F	0586
PIN8	65	M	0586
PIN9	71	F	0586

Gainfully employed

Person id	Industry	Salary	WorkMun.
PIN2	G	52 000	0586
PIN3	G	287 000	0580
PIN4	A	193 000	0586
PIN6	D	291 000	0586
PIN7	D	314 000	0580

Students

Person id	Educ.Type	Full/Part-time	StudMun.
PIN1	AdultEduc	100	0586
PIN2	Univ	100	0580
PIN5	Univ	100	0580

Statistical processing for different aims can be carried out with the above database. Chart 1.9 shows an example of two different data matrices that could be created from this database. These data matrices can be created and stored physically, but they can also be formed temporarily during processing. A conceptual data matrix focuses on the specific problem to be solved, i.e. the statistical analysis to be carried out.

Chart 1.9 shows data matrices for employment and commuting in the way that statisticians require for their work. The data matrices in the chart would be called *denormalised* in database terminology, as several cells are missing values.

Chart 1.9 Two data matrices for different statistical purposes

Data matrix: Employment register

Person	Age	Sex	Employed	Industry	Salary
PIN1	20	F	No	null	0
PIN2	23	M	Yes	G	52 000
PIN3	31	M	Yes	G	287 000
PIN4	32	F	Yes	A	193 000
PIN5	33	M	No	null	0
PIN6	40	F	Yes	D	291 000
PIN7	59	F	Yes	D	314 000
PIN8	65	M	No	null	0
PIN9	71	F	No	null	0

Data matrix: Commuting register

Person	ResMun.	WorkMun.	StudMun.	Commuting
PIN1	0586	null	0586	0
PIN2	0586	0586	0580	1
PIN3	0586	0580	null	1
PIN4	0586	0586	null	0
PIN5	0586	null	0580	1
PIN6	0586	0586	null	0
PIN7	0586	0580	null	1
PIN8	0586	null	null	0
PIN9	0586	null	null	0

1.4.3 What is a register survey?

The original data are generated in public administrative systems. Definitions of object sets, objects and variables are adapted to administrative purposes.

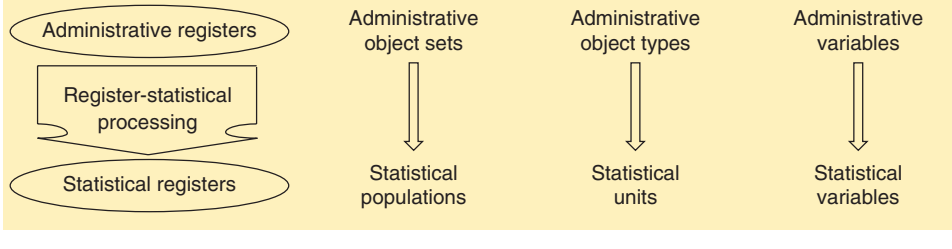
Every authority carries out controls, corrections and other processing suited to their administrative aims.

When an authority delivers *microdata* to a national statistical office, further selections and processing may be carried out by the authority to meet the needs of the statistical office. The authorities also have *metadata* as definitions, administrative rules and information regarding quality, which are based on the administrative authority's experiences and investigations. This information is important for those receiving the data at the statistical office.

Producing statistics directly from the received administrative registers is generally not a good idea because these are not adapted to statistical requirements. The object sets, object definitions and variables must be edited. And as a rule, some processing will be necessary so that the register fulfils the statistical requirements for population, units 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 statistical methods.

Charts 1.10 and 1.11 show three important components of this work. We have found that people tend to use administrative concepts as they are. This can be acceptable in some cases – but in other cases it is unacceptable. The three issues of how to define population, units and variables of a statistical register are important for the quality of the statistics to be produced.

Chart 1.10 From administrative registers to statistical registers – overview



A statistical population or administrative object set consists of *N objects* or *units* or *elements*. Of these three synonyms, we as a rule use the term *object* for the units in an administrative object set and the term *statistical unit* for the units in a statistical population.

Chart 1.11 From administrative registers to statistical registers – register processing

Administrative object sets		Administrative object types		Administrative variables	
	Matching object sets		Editing to find errors in objects and false matches		Standardising variable formats
	Handling of non-match		Handling of missing objects		Handling missing values and editing of variable values
	Selection of objects		Handling doublets		Coding
Statistical populations	Processing of time references	Statistical units	Creating derived objects	Statistical variables	Creating derived variables

1.4.4 A register survey: The Income and Taxation Register

In this section we give an example of a register survey where Statistics Sweden created the population, statistical units and statistical variables so that the final statistical registers differ from the administrative sources used. About 40 administrative and statistical registers were used to create three statistical income registers. The register-statistical processing follows the steps in Chart 1.11.

How is the Income and Taxation Register created?

The Income and Taxation Register (I&T) is an important part of Statistics Sweden's register system. It describes income distribution and can be used for regional income statistics. It is also the basis for longitudinal income registers used by university researchers and a micro simulation model used by the Ministry of Finance.

This register utilises many administrative sources, and many administrative variables are used to create important statistical variables. In addition to these administrative sources, the register system at Statistics Sweden must also be used. 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 register system to the Income and Taxation Register.

1. *Data generation at the National Tax Agency*
   The annual income self-assessment is based on tax returns from income earners and taxation decisions of the local tax authority. Both the income earner and the tax authority use statements of earnings for salary, sickness benefits and interest payments, which are the responsibility of employers, social insurance offices and finance companies. The National Tax Agency ultimately compiles this information. Tax returns, statements of earnings and taxation decisions can be changed and supplemented. Thus, data for one person can be very complex.
2. *Microdata deliveries to the Income and Taxation Register*
   The Swedish National Tax Agency annually creates databases that contain information on Sweden's population. The data files for one year – containing around nine million observations, each with around 300 variables – are delivered to the unit for income statistics at Statistics Sweden.
3. *Metadata to the Income and Taxation Register*
   Variable descriptions with names and definitions of variables accompany the deliveries from the National Tax Agency. Tax return forms, statements of earnings, taxation decisions and tax return instructions are also necessary for the correct interpretation of the data.
4. *Editing of data*
   The I&T Register receives data from many different suppliers outside and inside Statistics Sweden. Data from external sources are edited. Data from other Statistics Sweden registers have already been edited. Contacts between staff members at Statistics Sweden and persons at the National Tax Agency are important to obtain knowledge of changes in the administrative system. These in turn are important to ensure the quality of the register statistics – administrative changes should not be interpreted as actual income changes.

5. *Matching and selections*
   Many registers should be processed to create the different sub-registers included in the Income and Taxation Register. Observations from different sources are matched using Personal Identification Numbers (PINs). 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 compiled. One type of processing is to select persons aged 16 and older who were also part of the population on 31 December.
6. *Derived objects are created*
   A reference variable in Statistics Sweden's Population Register links each person to the *dwelling* where the person is registered as permanently living. All persons registered as permanently living in the same dwelling are classified as one *dwelling household*. These derived household units are created in the I&T Register and are used to produce income statistics for households.
7. *Derived variables are created*
   Many derived income variables are formed. For instance, the wage or salary amounts are aggregated from the different earnings data to become a person's *income from work*. Every person's total income from work and capital plus transfer payments minus taxes becomes the person's *disposable income*. Variables are formed for households, such as *household type, number of consumption units* and *disposable income*.

Chart 1.12 Different data sources for the Income and Taxation Register (I &T)

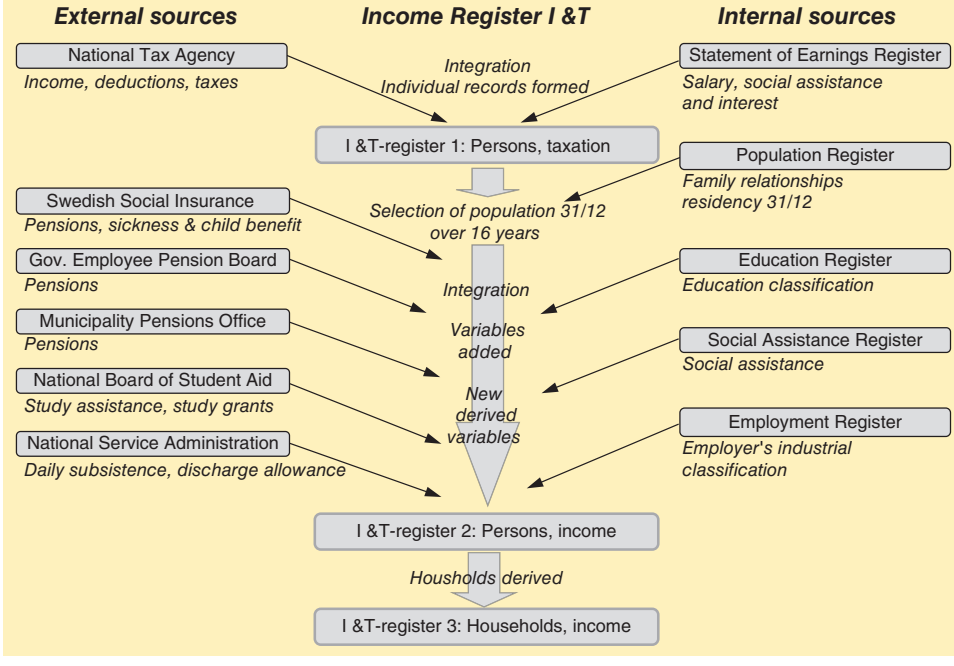


Chart 1.12 shows how the Income and Taxation Register receives administrative data from a variety of external sources and some Statistics Sweden

registers. The middle column shows the different phases when the sources are used in creating the new statistical register.

This example shows the importance of the principles in Chart 1.11. 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.

1.5 New demands and opportunities require new methods

The civic registers in the Nordic countries contain information on where persons and households have their permanent residence. This information is updated when persons change residence and household dwellings change. Thus, the administrative system generates daily data corresponding to the short form of the census.

This opens up new opportunities for timely population statistics and new kinds of social statistics. Longitudinal statistics based on these civic registers will also be something completely new for the national statistical office (NSO) – the fact that administrative systems cover all time provides new opportunities.

1.5.1 A new paradigm is necessary

The use of administrative systems instead of classical censuses and sample surveys represents a dramatic change for an NSO. Old competence and methods become obsolete and new methods and competence must be developed. This results in shifts in perspectives that translate into a new statistical paradigm.

We came to Statistics Sweden during 1989, just before the last classical census 1990. We noticed that several competent statisticians were negative about the change. The census was 'real statistical methodology' and the quality survey regarding the census was 'real quality assessment'. And all this would be lost when registers gradually replace the census!

Not only were some people sceptical about replacing the census with registers, data from tax forms were also viewed with scepticism. Could VAT tax forms be as good as our turnover survey? After careful analysis, it was admitted that there were only small differences between tax forms and the traditional survey.

New demands require a new methodological paradigm. Chart 1.15 in Section 1.5.3 presents four principles related to the new paradigm that are fundamentally important when an NSO works with administrative registers.

New competence

The staff and the national statistical office must build up new competence. Years of experience are required to fully understand how administrative systems function and how the data in the administrative registers should be understood. It is also important to spend time on establishing good contacts with the staff responsible for the administrative systems at the administrative authorities.

Managers must support this work of developing new competence and promoting discussions and seminars where the new sources and methods are discussed. To ensure that the administrative data are interpreted correctly, data

should be analysed and compared with other sources. Staff working with register surveys should spend time on this kind of editing. Discussion of the analytical results is important so that the level of competence in interpreting administrative data improves for the whole staff.

1.5.2 New statistical methods

A paper by Hand (2018) was read before the Royal Statistical Society, where Hand discusses what should be included in a generally accepted theory for the analysis of administrative data. Quality, coverage, relevance, record linkage and protection of confidentiality are challenges mentioned by Hand, and these issues are discussed in the following chapters. Hand also points out that statistical textbooks often only mention random sampling and sampling errors. Since administrative data are not created by random sampling, students will obtain a limited view of statistics. Hand (2018) discusses the following challenges:

Data quality

Challenge 1. Statistics teaching should cover data quality issues.
Challenge 2. Develop detectors for particular quality issues.
Challenge 3. Construct quality metrics and quality scorecards for data sets.
Challenge 4. Audit data sources for quality.
Challenge 5. Be aware of time series discontinuities arising from changing definitions.
Challenge 6. Evaluate the impact of data quality on statistical conclusions.

'Data = all'?

Challenge 7. Explore potential sources of non-representativeness in the data.
Challenge 8. Develop and adopt tools for adjusting conclusions in the light of the data selection processes.

Answering the right question

Challenge 9. Explore how suitable the administrative data are for answering the questions. Identify their limitations and be wary of changes of definitions and data capture methods over time.

Challenge 10. Report changes and time series with appropriate measures of uncertainty, so that both the statistical and the substantive significance of changes can be evaluated. The measures of uncertainty should include all sources of uncertainty which can be identified.

Causality and intervention

Challenge 11. Be aware that administrative data are observational data, and exercise due caution about claiming causal links.

Combining data from different sources

Challenge 12. Be aware of the risks that are associated with linked data sets and the potential effect on the accuracy and validity of any conclusions. Recognise that quality issues of individual databases may propagate and amplify in linked data. Develop better measures of overall combined data quality.

Challenge 13. Continue to develop statistically principled and sound methods for record linkage and evidence assimilation, especially from non-structured data and data of different modes.

Challenge 14. Develop improved methods for data triangulation, combining different sources and types of data to yield improved estimates.

Confidentiality, privacy and anonymisation

Challenge 15. Continue to explore anonymisation and deidentification methods.

Regarding Hand's Challenges 12 and 13, our experience from Statistics Sweden tells us that there are no risks of linking data sets with data on persons if the identity numbers are of Nordic quality. The best method to develop sound methods for record linkage is to improve the system of national identity numbers – probabilistic record linkage is not the solution. We return to this topic in Chapter 4.

New and old paradigms

Sample surveys are based on methods derived from an established theory: *sampling theory*. This theory has been developed within the academic world and statistical offices and consists of generally well-known terms and principles. Scientific literature and journals develop and spread the methodologies for sampling and estimation. Because the terms and principles are well known, people 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 and housing censuses. Measurement errors, design of questionnaires, coverage errors and nonresponse are methodological issues that also apply to 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.

Although register-based statistics are a common form of statistics used for official statistics and business reports, no well-established theory in the field exists. There are no recognised terms or principles, which makes the development of register-based statistics and register-statistical methodology all the more difficult. As a consequence, *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 surveys is not included in academic statistics. Statistical theory within statistical science is understood as consisting of probability theory and statistical inference. Sampling theory is included within this theoretical school of thought, but register 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 often carry out separate surveys. The building of *statistical information systems* is more common, and these constantly generate new data. A statistical theory is necessary to describe the general principles and to develop the conceptual apparatus for such statistical information systems. Register surveys should be included in this theory.

When the system approach is missing

Chart 1.13 illustrates a common problem. Estimates from four different surveys are compared, and these comparisons show clearly that the system approach is often missing in the work with statistical surveys. People are fully occupied with their own surveys and different surveys are published at different points in time. As a rule, most estimates are unique for one survey. However,

in Chart 1.13 we found one identical variable and created the table with corresponding estimates from each survey.

If we look at each survey separately, we do not see any errors except for the sample survey in (4) where we have margins for the sampling error. But when we look at the four surveys together, we understand that there must be more serious errors in these surveys. We thus need *a theory for systems of surveys and new methods for quality assessment*.

Chart 1.13 Employees by economic activity, November 2004, thousands

	Business Register		Employment Register	Labour Force Survey	
	Enterprises	Establishments			*Error margin*
Economic activity	***(1)***	***(2)***	***(3)***	***(4)***	***(5)***
Agriculture, forestry, fishing	35	37	37	26	*5*
Mining, quarrying, manufacturing	688	636	717	640	*23*
Electricity, gas, and water	21	22	28	29	*5*
Real estate, business activities	457	524	457	470	*20*
…	…	…	…	…	…
Education	382	408	431	462	*20*
Health and social work	836	684	675	675	*24*
Other service activities	142	163	175	168	*13*
Unknown activity	0	0	38	4	
Total	**3 763**	**3 763**	**3 924**	**3 778**	***43***

Why are there such large differences between the surveys? The estimates for mining, quarrying and manufacturing can be 636 000 or 717 000 – the inconsistencies are more serious than the sampling error.

The methodological work should have included the following three steps: (1) compare surveys and find errors and inconsistencies; (2) find out why we have these inconsistencies; and (3) reduce the errors and inconsistencies.

Chart 1.13 also illustrates that we only have one established way of giving a numerical description of the quality of published estimates – margins for the sampling error. There is no commonly used way of describing the quality of register-based statistics. However, the non-sampling errors of sample surveys are as a rule not described in the same clear manner as the sampling errors. We also lack methods here for giving a numerical description of the quality of published estimates.

The system approach – survey system design

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. The book was the first attempt to create a theory for register-based statistics and describe the methods used.

Statistics Denmark developed a system of coordinated and consistent register surveys to enable replacement of the traditional census with a register-based census. Hence, an entire system of register surveys was developed instead of creating a number of independently developed surveys. Statistics Denmark used the traditional census as a model and created registers that could produce population statistics, household statistics, employment statistics, education statistics and income statistics equivalent to the corresponding census statistics.

This was probably the first example of survey system design. The pioneer work done by Statistics Denmark constitutes a milestone in survey methodology.

To understand the significance of Statistics Denmark's achievement, Chart 1.13, which illustrates the lack of a theoretical approach, should be compared with Chart 1.14 based on the new register-statistical theory. Several registers were developed to produce yearly census-like statistics for small areas and small categories.

Chart 1.14 Register-based census statistics for one small municipality in Sweden

Population Register		Employment Register		Education Register					Income Register			
Age	Number of persons	Em-ployed	Not em-ployed	Com-pulsory	Upper secon-dary	Post secon-dary	Post-graduate	Not known	Yearly earned, SEK thousands *(1 USD ≈ 8 SEK)* 0	1-139	140-279	280-
0–15	1 358	-	-	-	-	-	-	-	-	-	-	-
16–19	384	84	300	297	67	0	0	20	120	259	5	0
20–24	328	197	131	55	233	31	0	9	38	153	105	32
25–34	719	622	97	88	423	201	1	6	15	141	320	243
35–44	962	846	116	83	554	319	2	4	19	106	381	456
45–54	910	776	134	171	500	226	4	9	24	89	363	434
55–64	1 071	791	280	295	491	271	7	7	21	145	468	437
65+	1 402	138	1 264	758	438	194	8	4	1	508	719	174

1.5.3 The basic principles of register-based statistics

We have formulated four basic principles for using administrative registers listed in Chart 1.15. We use these principles in the book and gradually introduce the register-statistical terms necessary for the discussions.

In a traditional statistical system with censuses and sample surveys, data are collected for each survey and all variables needed for each survey are simultaneously collected. This means that the established tradition is: *one* survey with *one* data collection gives *one* set of data for *one* specific purpose. Population and variables have been determined for this specific purpose.

Chart 1.12 with the Income and Taxation Register shows that none of this needs to apply to a register survey. Many sources are combined to achieve good coverage and to include all kinds of income. With these sources, three statistical registers are created for different purposes. According to the *Transformation principle*, all relevant sources should be combined when a

statistical register is created, and population and variables in this register may differ from the sources.

Chart 1.15 Four principles for using administrative registers for statistics

1. ***Transformation principle***
   Administrative registers should be transformed into statistical registers. All relevant sources should be used and combined during this transformation.
2. ***System principle***
   All statistical registers should be included in a coordinated register system. This system will ensure that all data can be integrated and used effectively.
3. ***Consistency principle***
   Consistency regarding populations and variables is necessary for the coherence of estimates from different register surveys.
4. ***Quality principle***
   The register system should be used for quality assessment of statistical surveys based on microdata comparisons with other surveys in the production system.

The *System principle* emphasises the importance of the register system as an important source, but registers in the system should follow the *Consistency principle* to avoid conflicting information from the different registers. The work with registers must be coordinated to achieve consistency. Populations and variables must be defined to suit the system, not individual surveys.

The *Quality principle* stresses that quality assessment must be done in a new way. In sample surveys, we describe the quality by computing confidence intervals. But how should we work with quality in register surveys? The quality issue is discussed in Chapters 5 and 6.

The goal: A coordinated statistical system

All microdata can be integrated in a national statistical system with a coordinated register system. In a coordinated system, the populations in the different registers are consistent regarding coverage, and the variables in different registers do not have contradicting values.

Of course, perfect consistency is difficult to obtain, but if coverage problems and variable inconsistencies are tolerable, we can say that the system is coordinated. Integration of microdata will become possible without disturbing problems regarding lack of consistency between populations and variables.

1.6 Preconditions for register-based statistics

Preconditions differ between countries for sample surveys, censuses and register surveys. Hence, the preconditions for statistical methods are different. The choice between cluster sampling and one-stage sampling depends on whether there is a Population Register or if you must use maps and address lists. Regression estimation and calibration are methods that depend on the number and quality of available register variables. This means that an *increased use of administrative registers will change the preconditions for all kinds of surveys*.

For register surveys, the differences between countries are even more significant. Legislation on national registration and the taxation of persons and enterprises determines the character of the administrative systems used in each country. The legislation regarding statistical production and protection of statistical data also differs. Therefore certain methodological issues are important in some countries but not in others.

The two main preconditions for using administrative registers for statistical purposes are stated in Chart 1.16.

Chart 1.16 Two preconditions for using administrative registers for statistics

5. ***Identity number principle***
   Unified systems of identity numbers are used in all administrative systems. The same identity number should follow an object over its lifetime.
6. ***Legal principle***
   A statistical office should have access to administrative registers kept by public authorities. This right should be supported by law as should the protection of privacy.

1.6.1 Reliable administrative systems

Reliable administrative systems will generate data of good administrative quality. Good administrative quality is a necessary but not sufficient condition for good statistical quality. The systems for tax administration and welfare programmes will gradually develop and change, and these changes will determine what administrative data can be used for future statistical purposes. Therefore, maintaining close and long-term relations with administrative authorities and politicians is important for national statistical offices.

The long-term strategy requires high-level contacts to promote strategic changes that will improve statistics production. The statistical office must explain to the administrative authorities how their data are used for statistical purposes. The statistical office also needs detailed information on how the administrative systems are organised and what changes are planned. Close and long-term contacts at all levels are required for these purposes.

What aspects of national administrative systems are important for statistical offices? We note two such aspects here, coverage and identity codes.

Coverage – the systems should cover all

The Nordic systems for child benefits are good examples of coverage. All children in defined age groups are entitled to a sum of money. All parents want the entitlement – but to receive the money, the parents must be registered as parents to the child in question and national identity numbers are required for the parents and the child. The residential address in the administrative register must also be correct so that the parents can receive the money. This system covers all children and all parents. All persons in the country will gradually be covered as the information in the system's registers is maintained and updated. The register will contain administrative, but also statistically important, links between all parents and children.

Good coverage requires that the administrative systems cover both urban and rural populations, rich and poor citizens, and small and big enterprises. The ideal is no selectivity. If suitable methods are not developed, selectivity will result in biased statistical estimates. For instance, in the Nordic countries all seriously ill persons will see a doctor, and all doctors know that cancer patients should be reported to the National Cancer Register. In this way we can be almost certain that all patients with a cancer diagnosis are in the Cancer Register. If rural or poor persons are underrepresented, estimated cancer incidence and mortality figures would have low quality.

Unified systems of identity codes

Identities are important in administrative systems. Legally important relations between persons, such as husband and wife, or parents and children, are registered with the identities of the persons in question. In many registers, the legally important relations between owners and different kinds of property are recorded with both the identities of the owners and the property. For taxpayers, it is important that the tax paid is recorded together with the identity of the taxpayer. Therefore, it is in the interest of each taxpayer to use a correct identity in each transaction. The legal importance of identities explains why identity data as a rule are of high quality in many administrative registers.

The best way to handle identities in administrative systems is to use *national identity numbers*. Persons, enterprises and property should be assigned unique identity numbers that are used in all administrative systems in the country. The same number should follow each person, enterprise or property over its lifetime.

Not only will administration become efficient; the statistical production system will become efficient. When administrative data are used for statistical purposes, it will be possible to link observations and create important statistical comparisons. Record linkage will be easy with unique national identity numbers, and the risk of false matches and false non-matches will be low. The statistical possibilities that national identity numbers create are explained in the following chapters.

It is advantageous if the identity numbers have no relation to any attributes of the objects that are to be identified. For example, identity numbers for persons should not depend on name, sex, or address of the persons, because such attributes can change over time. We use throughout the abbreviation PIN for national identity numbers for persons and BIN for national identity numbers for legal units representing enterprises.

1.6.2 Legal base and public approval

There are preconditions concerning legal base and public approval that enable the efficient use of administrative registers for statistics. These preconditions are discussed in UN/ECE (2007) and we build on that discussion here.

Chapter 2 describes the transition from a traditional statistical system into a register-based system. An important part of the work with this transition is to improve the preconditions for a register-based system. The Nordic countries have spent several years working hard to improve these preconditions.

Legislation determines what data are generated

The national administrative systems for taxation and welfare are based on legislation that determines the kind of administrative data generated within these systems. For example, if citizens pay income tax to municipalities, then the authorities must know where each citizen lives. The municipal taxation and welfare systems are the legal base for the Nordic administrative population registers. They are used not only for taxation and municipal welfare, but also for elections where the population register defines where each voter votes. For statistical purposes, this creates very good links between persons and geography that facilitate regional statistics. The administrative registers are updated every day, which makes possible timely monthly demographic statistics.

Legislation to improve the national statistical system

Politicians want to reduce the response burden of persons and enterprises as well as the direct costs for the production of official statistics.

- Legislation should provide the national statistical offices access to administrative microdata, including identities, and the right to use the data for official statistics and research.
- Legislation should provide statistical offices the authority to match data from different sources, and use data that were not originally generated for statistical purposes.
- Legislation could also instruct statistical offices to use data from administrative registers first, and to conduct sample surveys or censuses only if available administrative data are insufficient.
- Some laws have the sole purpose of making register-based housing and population censuses possible. For example, the Nordic parliaments have decided that all employers must provide information on where all employees work – the establishment address for all. This information is provided by income statements with data on employer identity, establishment identity, employee identity and wages, preliminary tax and social insurance paid. These income statements play an important role in the Nordic statistical systems, since we obtain important links between three different object types. The parliaments have also decided that all persons should be registered at the dwelling where they live. Then it will be possible to create statistics for households defined by the common dwelling in the register-based census.

Legislation on data protection

According to the second precondition in Chart 1.16, a national statistical office should have access to administrative registers kept by public authorities. This right, as well as the protection of privacy, should be supported by law. Improved legislation will take time – in the short run, agreements with ministries and public authorities will be necessary.

Legislation that gives a statistical office access to administrative data is discussed above, and the protection of privacy and confidentiality are discussed below.

The principle of *one-way traffic* is important for data protection. Microdata can go from administrative authorities to the statistical office but never in the reverse direction. The legislation on data protection should rest on a reasonable balance between protection of integrity on the one hand and increased costs and difficulties for statistics production on the other. An important task for top management at a national statistical office is to explain the consequences generated by proposed legislation to lawyers and politicians.

Public approval

The cooperation between register authorities and national statistical offices should be open and transparent. The fact that administrative data are used for statistical purposes should not be kept quiet; instead, the benefits of using administrative data and the efforts to protect integrity should be explained in open discussion and public debate.

It is important to explain that individual observations regarding persons are anonymous in statistics production, which is in contrast to how administrative authorities handle the same data.

If the national statistical office has a good reputation as trustworthy, gaining access to administrative data for statistics production will be easier. However, one mistake in the protection of integrity can immediately destroy this reputation.

Persons and enterprises do not want to be required to report to both an administrative authority and the national statistical office. Not having to do so will make public opinion more favourable to the use of administrative data for statistical purposes. The double provision of data becomes difficult to motivate – why respond to a questionnaire on the enterprise's turnover when you also submit a value-added tax return to the Tax Agency that includes the same information?

1.6.3 Political support to strengthen the statistical system

Political support and consensus is crucial for developing and strengthening the national statistical system. The national statistical office must have a role of coordinating and regulating the country's official statistics. This role must be recognised by ministries and other authorities that belong to the statistical

system. The following quotation from Dargent et al. (2018) is a reminder that this is not always the case.

> 'It is therefore noteworthy that incentives for strengthening statistical capacity are not always aligned within government or, in other words, there are trade-offs affecting the development of statistics policies. For example, governments need data for better decision making, but these same data can be used as a tool for citizens to demand accountability from governments for their decisions and thereby limit their discretionality. Consequently, the best and most transparent statistics policy can sometimes run counter the government's interests, creating incentives to ensure that these offices remain weak.'
>
> *Source*: From Dargent et al. (2018) Who wants to know? The Political Economy of Statistical Capacity in Latin America. Inter-American Development Bank.