Särtryck ur Statistisk Tidskrift 1966: 2

A statistical file system[1]

by Svein Nordbotten

1. Introduction

This paper is based on the assumption that the society shall continue to grow both in size and complexity. The condition for growth in a wanted direction is, however, that the decision-making units will at any time have access to an increasing amount of information describing the different aspects of the structure and changes of the society. This seems to represent an extremely requiring task for the statisticians.

The statisticians have so far been able to satisfy reasonable requirements partly by utilizing electronic data processing equipment. The development of these technical tools is probably at least as fast as the general development of the society and if we accept the electronic data processing systems as tools we want to utilize, the chances of being able to keep up with the increasing requirements seem to be favorable.

2. Statistical systems

By the statistical *product* we mean the total result of data collection and data processing in a period. If we assume that any statistical measurement may be affected by errors and that the results therefore are no more than estimates, the size of the product can be considered to be determined by the extension of the mass measured, the time needed to prepare the statistics, the number of statistics computed and the accuracy of the statistics. It has previously been usual to conceive the statistical products of the different fields such as population statistics, social statistics, financial statistics *etc.* In the future it may be more interesting and useful to discuss the result of the collecting activity or the *collection product,* as distinguished from the result of the estimation process or the *processing product.*

Let us consider the development of these two products. If we indicate the collection product by the number of individual questionnaires collected and the processing product by the number of pages of statistical tables printed, we find roughly that in Norway the collection product increased with 80 per cent and the processing product with 75 per cent in the last ten years. For both these products, the production time was about constant while the accuracy has increased during this period. Another feature of the development which the figures above do not reflect

[1] This paper is a translation of parts of the paper, Elektronmaskinene og statistikkens framtidige utforming, presented for the 7th Meeting of Nordic Statisticians in Helsingfors 1960. The original paper also contained considerations about future developments in electronic data processing equipment. In contrast to the statistical system, the technical development considered has already been realized. The technical parts of the paper are therefore absolute in this connection and not included in the translation.

is the periodicity of the statistical activities. Traditionally the month, quarter and year are chosen as periods for statistical observation. The tendency is supported strongly by the decennial censuses which often are concentrated at the end or beginning of a decennium. This tends to a high collection activity in the first part of the month, the year and the decennium with a lagged processing activity following.

Today's strong tendencies toward expansion and differentation in economic activities will probably require a development of the statistical activities at the same rate in the next 10 years to come. If the statistical system remains unchanged, we shall in Norway in 1970 have collection and processing products about three times the size of those we had in 1950. With constant production time this may give even more marked variation in the activity.

If the society is going to get full use of the statistical product, the production time must be shortened, perhaps to a certain degree at the cost of accuracy. To-day users of statistics must often make their decisions by extrapolating from statistical results which refer to a period or point of observation which in some cases may be several years behind. When the society grows more complex and changes faster, the risk involved in the use of old statistics becomes greater. The risk by extrapolating from relatively accurate, but old statistics, must therefore be compared with the risk by using up-to-date, but statistically more uncertain estimates. There has been a marked tendency to choose the latter alternative, for example by increased use of sample surveyes in the post-war period. If the development is going to require more up-to-date statistics, this will imply that the variations in the activity will become even larger, *i.e.* short periods with very intensive data collection and processing with intermediate periods with low utilization of the capacity.

The relative long production time of the existing system implies that the statisticians must plan rather general purpose results in order to cover as many as possible of the requests which may occur. It may be discussed how efficient such general purpose plans really are. With a shorter production time it might be efficient to pay more attention to individual requests when they are made and repeat processing when necessary. That means, we should keep the data in a state of *statistical readiness.* The statistical production would therefore be oriented towards fast service of individual needs rather than the time-consuming preparation of large general purpose statistical volumes and the most frequently requested information must of course still be printed in general survey volumes.

The problem due to the increased requirements to the statistical product, may perhaps be solved by supplementing the present processing with a *statistical file system* (statistical archives). By a statistical file system is meant a system in which the data collection is done continuously and independently of the traditional statistical fields and observation times, and in which the processing of filed data is carried out when needs occur.

3. Continuous data collection

Let us consider how such a statistical file system works, and we start with data collection. In order to obtain an up-to-date statistical file, collection, editing, and filing should be more continuous instead of the present periodical pattern. This may not necessarily imply larger work loads for the respondents or more inaccurate results. The following simplified example illustrates the idea of continuous collection. The practice of collecting data for an annual survey is to collect all the material once each year from all units. In a continuous scheme the units are divided at random into for example four groups. Each group reports at the end of different quarters on their activities in the last four quarters. At the end of each quarter we shall have a sample for making estimates for any previous period at the same time as complete data exist for any period older than one year.

The continuous collection will no doubt require extensive use of sampling methods and a modern field organization. Today data for the survey of manufacturing and the wage statistics are collected separately. It will probably in many cases be more efficient to integrate the collection, *i.e.* collect as much as possible from a respondent when we have made a contact instead of repeative visits or requests.

When statistical classification standards become more established, the questionnaires may to a great extent be precoded, or the respondents themselves may carry out essential parts of the coding. This may lead to great savings as we shall see later.

4. Statistical data files and registers

How should the data be filed? The data may be divided into three groups: The *active files,* the *historic files* and the *statistical registers.* In the active files the most recent observations are stored while the historic files are used for all historic data. The latter files will grow and the problem of optimum amputation will therefore have to be considered. The statistical registers will be used as a basis for collection and will contain information such as the identification numbers, names and addresses of the units.

A very important question is how the files should be organized. The data are distributed to the files according to the statistical unit to which they refer. In each file the main order will probably be set up by the identification numbers of the units. In connection with the identification it is extremely important to emphasize that the files must contain cross references to other files. All data related to a unit including identification can be considered as a record. Within the record the ordering of data may be cronological. Each item consists of two parts, the identification of the characteristic observed and of time, and the value of the observation. The concrete design of these records will of course depend to a great extent on the storage medium available.

When new data are collected, certain filing routines have to be performed. The new data must be compared with data in the files and an editing carried out. A check will be done to see whether the unit is already registered, whether the characteristic has been observed previously, whether there is any change in the value of the characteristic *etc*. Because of limited storage capacity the unchanged value of a characteristic may not be recorded. A validity control according to specified rules must also be integrated in the filing routines. If some data are rejected by the control, they will still be filed as preliminary data because, even erroneous, they represent the best we have, but a message is produced about the rejection. When corrected data become available, they will substitute the preliminary data in the active file and will also be transferred to the historic file. The last possible observation of a unit is its death. The unit will then be excluded from the register and the active file and an end mark is written in the historic file. The organization of the two files may be the same in other respects.

5. Data processing and estimation

Today we have in general to postpone the processing until the collection is completed. The purpose of the file system is that the statistician at any time, will get the most recent data from the active file for the analysis of a given problem. The historic files are on the other hand the source for those interested in the development over time. From these files the development can be studied at an *individual basis* in contrast to the analyses based on aggregates for traditional groups which so far have been the most usual.

Another point which we shall discuss is the following. Owing to the long production time of today, relatively high overhead costs, infrequent surveys and the fact that collection is closely connected with the following processing, it is essential that information requests anticipated but not received must be taken into account when collecting and processing. A significant portion of the processing may result in statistics which never are used. In a statistical file system the variable costs per unit, statistic or table will be more dominant because the data already are filed and it will therefore be more efficient to satisfy special needs when they occur. The files will probably encourage to make experiments as for example in the field of sequence-analytical estimation in order to reduce processing costs.

Because of the continuous collection and the size of the files, the estimation aspect will probably be relatively more important than it has been the case so far. This requires that the processing has to be organized for solving specified problems and that the problems have to be formulated more rigorously in accordance with the statistical theory.

So far a similar development as described has been impossible because convenient filing of data was very expensive and the size of the data mass was to a great extent the decisive factor in regard to the cost of processing. Therefore extensive tabulations have to be made when a statistical survey was taken. In a statistical file system with electronic data processing, the purpose is to beat this obstacle and it is believed that in the long run the costs of many special processings will be less than that of general processing for both known and unknown needs.

The purpose of the processing in a statistical file system is to produce up-to-date results from file-data which permit a maximum number of combinations. It is obvious that the system must permit parallel work on several tasks. It must be possible to do continuous and parallel filing at the same time as several processing routines are carried out.

6. Conclusion

Compared with the traditional system, the statistical file system implies the possibilities of

— a smooth collection and efficient utilization of the capacity of the system,
— the utilization of data across traditional statistical fields,
— studies of the development over time on a unit basis, and
— statistical readiness for fast service to users of statistics from the data files.

KUNGL. BOKTR. STHLM 1966