

## Developing statistical theories for register-based statistics

Li-Chun Zhang  
STATISTISK SENTRALBYRÅ, NORGE

For some decades now administrative registers have been an important data source for official statistics alongside survey sampling and population census. Not only do they provide frames and valuable auxiliary information for sample surveys and censuses, systems of inter-linked *statistical* registers (i.e. registers for statistical uses) have been developed on the basis of various available administrative registers to produce a wide range of purely register-based statistics (e.g. Wallgren and Wallgren, 2007). A summary of the development of some of the key statistical registers in the Nordic countries is given in UNECE (2007). The next census in 2011 will be completely register-based in all Nordic countries. Reduction of response burden, long-term cost efficiency, as well as potential for detailed spatial-demographic and longitudinal statistics are some of the major advantages associated with the statistical use of administrative registers.

The trend towards more extensive use of registers is increasingly being recognized by statisticians around the world. However, also being noticed is a clear lack of *statistical* theories for assessing the uncertainty of register-based statistics. It is tempting to reflect over the historical development of survey sampling for comparison. The *representative method* was presented by Kiær at the ISI meeting in 1895. Despite that he was not able to defend the approach on a theoretical ground, the practice kept evolving in the subsequent years. In 1924, the ISI formed a committee to investigate the feasibility of the representative method. The reporter Jensen stated: "When ISI discussed the matter twentytwo years ago, it was the question of the recognition of the method in principle that claimed most interest. Now it is otherwise. I think I may venture to say that nowadays there is hardly one statistician, who in principle will contest the legitimacy of the representative method. Nevertheless, I believe that the representative method is capable of being used to a much greater extent than now is the case". Indeed, breakthroughs did not come until some 30 - 40 years after Kiær's initial presentation. Today, the contributions by Bowley and, in particular, Neyman (1934) are generally taken as the starting point of the theoretical development of the so-called design-based approach to survey sampling.

Register-based statistics, when viewed in a mirror of history, appear currently pretty much at a pre-Neyman stage in their theoretical development. We believe that the key issue here, from a statistical methodological point of view, is the *conceptualization* and *measurement* of the *statistical accuracy* in register data which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, as one is able to do in survey sampling.

Administrative registers certainly do not provide perfect statistical data. In order to identify the various potential error sources, we thought it may be helpful to chart the *life cycle* of statistical micro data in a way similar to what Groves *et al.* (2004, Figure 2.5) have done for sample survey data. The resulting two-phase diagram is shown in Figure 1.

The entire life cycle of statistical micro data is generally divided into two phases. The first phase concerns data from a single source, the second phase concerns the possible integration of data from different sources. The life cycle of sample survey data as charted by Groves *et al.*

falls in the first phase. But their concepts have been extended to accommodate data from the administrative sources as well. Let me elaborate along the line of representation. In survey sampling the target set would contain the units of the target statistical population, whereas the accessible set would correspond to the sampling frame. The difference between the two is known as frame error. Next, the accessed set would correspond to the gross sample, whereas the observed set would only contain the respondents. The selection error corresponds then to the sampling error, and so on. But the distinctions are equally applicable to an administrative data source, in which case the difference between the target and accessible sets of objects are often due to feasibility reasons. For instance, a job register, among others, may be intended to provide the basis for the administration of sickness and child-care benefits. The target set should ideally contain all jobs that are 'substantial' enough to qualify for the benefits. To facilitate the reporting in practice, however, only jobs of certain regularity are required to be registered, hence accessible at all. Next, the accessed set would contain all the jobs that are actually registered, and the validated set the ones that are deemed admissible. Inevitably, some selection errors are associated with the reporting/registration process, while some of the reported jobs may be inadmissible due to confusion surrounding the administrative routines or regulations.

The second phase of integrating data from different sources is an essential feature in register-based statistical production. The idea is to make secondary uses of data that have been collected for a different purpose primarily. Above all, integration with the so-called population base registers, such as the *Central Population Register* and the *Business Register*, is necessary in order to incorporate the register data into the statistical system at all, because the target statistical population and variables for integration are almost always different from those of the primary phases. Hence the necessary distinction between objects at the first phase and units at the second phase. In the example of the job register above, what are primarily registered are job events such as hiring and dismissing, which are the objects at the first phase. To turn this information into, say, employment status of persons, i.e. the units for a secondary statistical purpose, conversion from the primary objects must take place. It is important to emphasize that, from the same perspective, also sample survey data should be integrated into the statistical system in order to achieve greater secondary usability, by which the life cycle of sample data will be extended beyond the first phase.

A shared understanding of the potential error sources can help us to collocate and coordinate the different research efforts. Recently, several research initiatives have been undertaken at Statistics Norway. For instance, the uncertainty of the register-based census employment status is being studied. Comparisons are made to the Labor Force Survey data. Since a main cause of difference is the employment definition in each source, which is referred to as relevance error (2<sup>nd</sup> phase, Figure 1), a theory for valid and equivalent statistical micro data is being developed in connection. Another example is unit errors (2<sup>nd</sup> phase, Figure 1) in the household register. The statistical unit 'household' does not exist in any administrative source; it must be constructed from the relevant primary units, such as person, family and dwelling. Errors are unavoidable. Again, the problem is rooted in the secondary use of data. A statistical theory for household unit errors has been developed (Zhang, 2010), which allows one to assess not only the uncertainty in the household statistics, but also the statistics that are based on household units. Other topics that have been studied include a modeling approach to registration delays and mistakes (i.e. measurement error, 1<sup>st</sup> phase, Figure 1) for producing statistics at detailed levels (Zhang and Fosen, 2010), and an imputation-based simultaneous prediction approach (Zhang, 2009) for statistical registers (i.e. selection and missing data error, 1<sup>st</sup> phase, Figure 1). All these topics are highly relevant for the forthcoming register-based census.

To summarize, all national statistical institutes face currently the challenge of finding a way between budgetary constraints and ever increasing demand on statistical information. Efficient use of all data available is naturally an option that must be explored. The 20<sup>th</sup> century has witnessed the birth and maturing of sample surveys. The 21<sup>st</sup> century will be the age of data integration, and administrative register data are a major part of it. As statisticians, it is our duty to make it come out right.

## References

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2007). *Survey Methodology*. New York: Wiley.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, nr 97, 558-606.

UNECE (2007). *Register-based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics*. United Nations Publication, ISBN 978-92-1-116963-8.

Wallgren, A. and Wallgren, B. (2007). *Register-based Statistics - Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd.

Zhang, L.-C. (2009). A Triple-goal Imputation Method for Statistical Registers. Paper presented at the UNECE Work Session on Statistical Data Editing, Neuchatel.

Zhang, L.-C. (2010). A Unit-error Theory for Register-based Household Statistics. To appear in *Journal of Official Statistics*.

Zhang, L.-C. and Fosen, J. (2010). Assessment of Uncertainty in Register-based Small Area Means of a Binary Variable. Submitted.

**Figure 1.** Two-phase (primary and secondary) life cycle of statistical data from a quality perspective. Stage of data production (square frame) and type of potential errors (circle frame).

