Disclosure control on censuses and surveys. Basic principles on micro data protection in Statistics Finland.

Janika Konnu[1]

[1] Statistics Finland, Finland
e-mail: janika.konnu@stat.fi

Abstract

Statistical Offices collect huge amount of information on persons and enterprises. Statistics Finland gets most of the information from registers and that improves the quality of the data. It is only natural that researches want to use the data Statistical Offices have collected. The disclosure risks and the protection needed for the data differ greatly between the register based census surveys and sample surveys. In our paper we will describe some basic principles on protection and some differences in each case. As a rule much stronger Statistical Disclosure Control methods are needed when protecting a census survey than protecting a sample survey data.

1 Introduction

Statistics Finland is responsible for compiling most of the official statistics in Finland. That is the reason why Statistics Finland has access to many registers. Different data sets from registers and from researchers can be linked easily because of Finnish personal and business identity codes. All this information in Statistics Finland is very interesting for both native and foreign researchers and the amount of applications for micro data is increasing.

In Finland the Statistics Act (280/2004) restricts releasing personal data referred in the Personal Data Act (523/1999). This is interpreted so that personal data can be released for researching purposes only and even in this case the data must be in anonymised form. Here the anonymisation means that not only the direct identification but also indirect identification must be prevented. Direct identifiers in case of personal data are variables such as personal identity code, passport number, name, and address or phone number.

In case of business data, Statistics Act (280/2004) is interpreted so that data can be released for researching purposes if the data is anonymised. If the data cannot be anonymised, it can be used in the premises of Statistics Finland. For business data, variables like the name and the business identity code of the enterprise can lead to direct identification.

It is possible to identify an individual in a data set without any direct identifiers. This kind of identification is called indirect. Usually one of the key variables that lead to indirect identification is regional information. Data, which includes very detailed information on the area where enterprise is located or a person lives, can lead to identification. If the snooper (i.e. the person who is trying to identify some records in the data) has personal data in hand, in case of a small municipality, he/she only needs information on person's age, gender and occupation to identify a person. Even in bigger municipalities there are some rare occupations or combinations of the key variables mentioned earlier so that some of the people can be easily identified. Regional information is very sensitive when disclosure risks are measured. If we take a look on business data, the biggest companies are easy to identify using the information on profit etc. The smaller enterprises are always harder to identify, even with very detailed information because there are so many of them.

2 Disclosure risks

Disclosure risks can be defined for an individual record or for the whole data in hand. All the measures are based on some disclosure scenario. In the worst case scenario the snooper has some external data of the population and the data incluces identification codes and most important key variables and there is no error in his data. The worst case scenario also assumes that the snooper attempts to identify all records in the data. In more realistic scenarios only some key variables or some individuals are known by the snooper. Usually people know basic information about their relatives and neighbours and can use this information to identify some person. People that are commonly known like celebrities, athelitics and politician etc. have higher risk for disclosure. These different scenarios can be taken into account when measuring disclosure risks. As a rule, realeasing a census means high disclosure risk and to decrease this risk only samples of a census surveys are releashed (see chapter 2.2).

The actual measures of the disclosure risks however are normally used only for theorethical purposes and this is why we describe only the basic idea as it has been explained in Konnu (2006). When the data protector is trying to measure the disclosure risk, it is important to specify the key variables for disclosure. Regional information is always one of the most sensitive variables, but also variables discribing a person or an enterprice are sensitive. Data protector cross-tabulates contingency tables using these key variables and the *k*-th cell in the table corresponds the combination *k* of the key variables. The population *P* has a partition which is defined by combinations $\{1,...,k,...,K\}$. For each cell there are several individuals and let's denote F_k for the number of population units in each cell. For population *P* this

frequence is often unknown but for the sample *s*, which is realeashed, the corresponding frequency f_k is known. For a sample, f_k can be 0 and so only a subset of the combinations *K* might be included. For the measurements, let's propose that only the subset of combinations for which $f_k > 0$ are included.

Individual risk of disclosure for unit $i \in s$ can be definied as a probability:

$$\rho_i = \Pr(i \text{ correctly linked with } i^* \mid s, P, L_i) \Pr(L_i), \quad (1)$$

where $i^* \in P$ and the event L_i is that the snooper attempts to link the individual $i \in s$ and some individual in *P*. For simplicity we can assume the worst case scenario, which was described earlier, takes place and then the snooper attempts to indentify all the units in the sample *s*. Then the latter probability in (1) makes $\Pr(L_i) = 1$ for all *i* and we get

$$\rho_i \le r_i = \Pr(i \text{ correctly linked with } i^* \mid s, P, \text{worst case scenario}). \quad (2)$$

If we think about the information snooper have access to, he has the partition described earlier but no other information. Then there is no way to make difference between the individuals *i* in cell *k* and each unit of the F_k can be linked to any unit in f_k. Because of this, the disclosure risk r_i for an individual record *i* is the same for all units in cell *k*. Now we can define the disclosure risk as r_k for all *i* in cell *k*. If the frequency F_k is known the probability for indentification can be defined as $r_k = \frac{1}{F_k}$. In Statistics Finland the frequencies F_k are usually known but when they are not, this measure must be approximated. Many different approximations have been succested but as mentioned earlier, these measures are usually used for theoretical approaches only.

2.1 Sample surveys

Taking a sample over the population and then collecting the information forms a sample survey data. With every sample survey there is also some non-response involved. Usually snooper is trying to find some individual he already knows or spot some unique individual in the data and then try to identify it. In case of a sample survey, snooper can't be positive about the identification, because the person or enterprise he is interested might not be in the sample at all. Survey non-response is usually higher for the rare cases in the population and this is another reason why a sample survey data doesn't have such a high disclosure risk.

2.2 Census surveys

Census surveys are usually register-based data in Finland. Census surveys are sensitive for disclosure because they include all possible individuals and all snooper needs to do is spot the individual in his interest. In the worst case scenario described in the beginning of the chapter, it would be possible to match almost every unit in the data. That is the reason why in principle Statistics Finland never releases a census data but a sample of it (Statistics Finland, 2005, p. 2).

3 Data protection

Data must be protected against disclosure before it can be releashed even to researchers. In Finland the legistlation is quite strickt and not only direct but also indirect identification must be prevented. When data is releashed for reasearching purposes, the use of the data must also be taken into account before applying the protection. For that reason researchers must attach a research plan into the application for a data set. Even if the reseach plan is of a high quality, the data protector and the researcher must discuss about the suggested protection. In some rare cases it turns out that some variable, which is not so important for the research in the data protecter's point of view, is actually very important. When the protection is agreed by both sides, data can be protected and then releashed.

Data protecting prosess begins with removing all direct identifiers and after that the need for other protection must be assessed. Usually only those variables that can be used to identify some individual are protected. This means that information descriping a person or an enterprise must be protected but when survey includes information about opinions or way of behaiving, these variables usually don't need any protection. However this can't be taken as a rule. Variables like person's height or weight can descripe a person in detail.

Variables that include detailed information must be categorised and sometimes for categorical variables even broader categories must be formed. For countinuos variables it is possible to use rounding or categorisation when needed. If researcher needs detailed information on a continuos variable, also some noise can be applied to the values of the variable. For both type of variables the individuals that are easiest to identify are those that have very large or very small values. For categorical variables, this means that the first and the last category must be broader than others. For continuous variables, data protector must define a threshold below or above which the values aren't releashed to protect this kind of individuals. This means that the largest values are replaced by information that the value is the above some threshold

value and for the smallest that they are below the threshold. This method is called top and bottom coding and it is very commonly used in Statistics Finland.

The new SDC methods, that use probabilities and modelling to protect data, aren't in use in Statistics Finland. We feel that these new methods aren't tested broadly enough and some of them also mean extra work (and need of knowledge) for data user. We are currently testing these methods and thinking about their usability, but they probably won't be in use in the near future. Some of those methods have potential for protecting Public Use Files (PUF) which are data set available for any citizen in a country. In Finland however the legislation forbids the use of PUFs.

3.1 Sample surveys

Sample survey data doesn't include the whole population and because of that, sample surveys are easier to protect than census surveys. Data protection for a sample survey data is usually just making sure the categorised variables aren't giving too detailed information and the values of the continuous variables aren't giving opportunity to match with some external data.

For categorised variables, data protector has to check the categories for at least variables of region, occupation etc. If data doesn't include many key variables, it is possible to give these variables in quite detailed level. But usually data has so many key variables that categories have to be broad to prevent a situation where a sample unique is actually a population unique too. So at least some tables must be compiled to check this situation.

For continuous variables the first thing is to prevent the direct matching with an external data. This means rounding the variable or adding noise to it. Usually continuous variables must be top coded too and top coding is used at least for variables on income and business profit etc.

3.2 Census surveys

In case of a census, whether it is personal or business data, disclosing at least some record in the data is not that hard. If the whole population is included, it doesn't really help that much if data protector removes those variables that lead to direct identification. When the snooper has some specific individual in his interest and some basic information about the target, it is easy to find it in the data. This is the reason why first of all a sample must be drawn from the census survey data. The method for sampling and the sampling ratio depends on the use and other protection on the data.

Basically when the sample of the census survey is defined, the protection is very similar to a sample survey. Same kind of combinations must be checked and same kind of protection must be applied. However the sample size for data to be released is usually higher in case of a census survey data than when the actual survey is based on the sample. And for a census survey, all the individuals can be in the sample, because there is no non-response. This means that the sample to be protected includes many of those easily identified rare cases. So even if the methods for protecting the data are similar, the need for categorising and broader categories is justified. Sometimes there is also a need to delete some records from the sample because they cannot be protected properly in other ways.

References

Konnu, J. (2006) *Mikroaineistojen tilastolliset tietosuojamenetelmät henkilötilastoissa. (Statistical Disclosure Control Methods for Personal Microdata; in Finnish only).* Master's theses in statistics. University of Jyväskylä, Department of Mathematics and Statistics, Jyväskylä.

Personal Data Act. 523/1999. Helsinki.

Statistics Act. 280/2004. Helsinki.

Statistics Finland. (2005) *Ohje käyttölupien myöntämisestä Tilastokeskuksen perusaineistoon. (Guidelines on granting permissions to use Statistics Finland data files; in Finnish only).* Reg. No. TK-00-128-05. Helsinki.