Working Paper No. 18 (Summary)

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (v): Risk assessment

SOME REMARKS ON THE INDIVIDUAL RISK METHODOLOGY

Invited paper

Submitted by the National Institute of Statistics (ISTAT), Italy [1]

[1] Prepared by Silvia Polettini (polettin@istat.it).

Some remarks on the individual risk methodology

Silvia Polettini (Istat, Rome Italy)

The paper discusses some aspects of the individual risk methodology, that was initially proposed by Benedetti and Franconi (1998). The original formulation defines a record-level measure of re-identification, called the *individual risk*, that can be estimated exploiting information on the sampling design. This methodology is currently implemented in the testing version of the software μ-Argus, developed under the European project CASC.

When dealing with social surveys, that is the context where the individual risk methodology is best suited, it is reasonable to hypothesise that identification, which consists of linking a sample unit to a population unit, is performed based on a set of known identifying or *key* categorical variables. This implies that the individual risk depends on the joint distribution of the key variables, e.g. on the size f_k, F_k of subgroups of units having a given *combination* of key variables in the sample and population, respectively. Unlike the approaches that propose a record-level measure of risk based on the concept of sample uniques (e.g. Skinner and Elliot, 2002), the risk is defined for any record in the sample. The measure also differs from those based on the sample frequency of combinations, because inference on the sizes F_k of population subgroups is performed. The method shares with the above mentioned strategies the inferential nature and the approach to protection, respectively. Indeed, having estimated the individual risk for each record in the sample, protection is ensured by applying local suppression to high risk individuals only.

The paper discusses the formalisation of the individual risk function for files of independent units. Upon defining the disclosure scenario, the individual risk measure is linked to the probability of re-identification of a single record given information on a set of key variables observed on the whole population. Based on such connection, an overall measure of risk, called the *re-identification rate*, is proposed. Although this is a measure at the file level like the ones discussed by Skinner and Elliot (2002), it exploits the probability of re-identification of each sampled record. In particular, it is defined in terms of the expected number of re-identifications in the file to be released. Whenever the individual risk methodology is used to protect a sample by local suppression, the user is requested to select a risk threshold that classifies individuals into safe or unsafe. The paper investigates how the re-identification rate may be exploited for selection of a proper risk threshold using a measure of target "safety" of the whole file.

As proposed by Benedetti and Franconi (1998), estimation of the individual risk relies on the expression $r_k = \sum_{h \geq f_k} \frac{1}{h} P(F_k = h \mid f_k)$. We notice that the latter expression identifies the individual risk with the risk of re-identification *from the Agency's viewpoint.*

According to the superpopulation approach introduced by the authors, $F_k \mid f_k$ is modelled as a negative binomial random variable with parameters f_k, and success probability p_k; this implies that the individual risk can be equivalently written as the order -1 moment of a negative binomial random variable:

$$r_k = E\left(F_k^{-1}; f_k, p_k\right) = \int_0^{\infty} \left(\frac{p_k e^{-t}}{1 - q_k e^{-t}} \right)^{f_k} dt$$

The paper provides theoretical results that enable rewriting the above formula in terms of the integral representation of the Gauss Hypergeometric series,

$$F(a,b;c,z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} \left(1 - t^{c-b-1}\right) (1 - tz)^{-a} \, dt \, .$$

The theory of Hypergeometric functions permits to derive alternative expressions that can be exploited for the purpose of estimating the individual risk; approximating formulae are also provided that hold for moderate to large f_k.

References

Benedetti, R. and Franconi, L. (1998): Statistical and technological solutions for controlled data dissemination, *Pre-proceedings of New Techniques and Technologies for Statistics*, 1, 225-232.

Skinner, C.J. and Elliot, M.J. (2002): A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society, Series B*, 64, 855-867