HOMEWORK 3

Ganna Fagerberg / Statistical methods

Problem 1

1.1. Define disclosure

In a broad sense, disclosure is the process of making some (confidential) information known. There is no universally agreed definition of disclosure as it largely context dependent. For example, Lambert (1993) argues that for some, disclosure occurs when the anonymity of a respondent (or, more generally, a unit) is violated, whereas for others it occurs when sensitive information is revealed from the released data, even if it is erroneous.

There are three major types of disclosure: (i) identity disclosure (or, re-identification), (ii) attribute disclosure, and (iii) inferential disclosure. Identity disclosure refers to the inadvertent release of a record that can be linked to a particular unit, e.g., through direct identifiers or through a combination of unique key characteristics. Attribute disclosure refers to the process of retrieving some new information about a unit or units with or without her/their prior re-identification. The information about inferential disclosure is given below (Benschop *et al.*, 2019).

1.2. What do we mean with inferential disclosure? Given an example

Inferential disclosure is a type of disclosure which enables an intruder to draw conclusions from the information retrieved from released data with a high level of accuracy. For example, if an intruder is interested in inferring information about a unit's income from the purchasing prices of homes, it might be possible due to usually high correlation of income and purchasing prices (OECD Glossary of Statistical Terms).

1.3. Why is uniqueness neither a sufficient nor a necessary condition for attribute disclosure? Illustrate with examples

First, let's define the term uniqueness. Uniqueness in general refers to some unique key characteristic of a unit, making the unit itself unique. Uniqueness can be either an attribute of a population, or of a sample. A released data that contain information with some key unique characteristics on a unit has a higher risk of re-identification.

However, it is neither a sufficient nor a necessary condition for disclosure. The reasons for this are the following. It is not sufficient since the unique unit must be represented in a sample in order to be identified and the intruder must be sure that the unit is indeed unique.

It is not necessary as retrieval of key information about a unit or units of interest may also occur in other ways. For example, one common way involves forming the so-called coalitions. Coalitions may occur if several units in a population share the same unique key characteristics. If there is an interest in disclosing information about a unit having these unique characteristics, other players with the same characteristic may form a coalition to freely exchange information between each other and to disclose information about a unit of interest left outside the coalition.

It can be also possible to disclose some sensitive information even when there are no unique key characteristics to rely upon. For example, if all units share the same key characteristics, it may still be possible to disclose some sensitive information for all the units without re-identification of individuals.

Also, re-identification can be possible if the intruder disposes of information that a particular unit has taken part in a survey and hence her information must be present in a released data (Carlson, 2002).

Problem 2

2.1 Use the data from Tables 2-5 to assess which cells are at risk of disclosure using the dominance rules (1,60) and (2,90) and the p%-rule (1,11)

The dominance rule (1,60)

TABLE 2	Region					
Industry code	North	East	West	South	Missing	Total
103	-	0	92 000	20 000	-	**112 000**
140	22 000	1 000	-	-	-	**23 000**
142	1 238 000	58 000	97 000	220 000	-	**1 613 000**
145	63 000	146 000	112 000	495 000	-	**816 000**
Total	**1 323 000**	**205 000**	**301 000**	**735 000**	**-**	**2 564 000**

The dominance rule (2,60)

TABLE 2	Region					
Industry code	North	East	West	South	Missing	Total
103	-	0	92 000	20 000	-	**112 000**
140	22 000	1 000	-	-	-	**23 000**
142	1 238 000	58 000	97 000	220 000	-	**1 613 000**
145	63 000	146 000	112 000	495 000	-	**816 000**
Total	**1 323 000**	**205 000**	**301 000**	**735 000**	**-**	**2 564 000**

The p%-rule (1,11)

TABLE 2	Region					
Industry code	North	East	West	South	Missing	Total
103	-	0	92 000	20 000	-	**112 000**
140	22 000	1 000	-	-	-	**23 000**
142	1 238 000	58 000	97 000	220 000	-	**1 613 000**
145	63 000	146 000	112 000	495 000	-	**816 000**
Total	**1 323 000**	**205 000**	**301 000**	**735 000**	**-**	**2 564 000**

2.2. Assume that you are contemplating using the two dominance rules (1,80) and (2,80). Explain why the first rule is redundant when you have the second rule.

In practice, a combination of dominance rules is often used in order to determine whether a cell is prone to disclosure. A cell is considered as safe if it is safe according to some carefully chosen two combination rules. For example, according to both $(n = 1, k = 60)$ and $(n = 2, k = 90)$ rules or according to $(n = 1, k = 80)$ and $(n = 2, k = 95)$ (see lecture materials). Note the following. According to these rules we check that the contribution of the largest contributor (as checked by, e.g., $(n = 1, k = 60)$ rule) and of the two largest contributors (as checked by, e.g., $(n = 2, k = 90)$ rule) simultaneously do not contribute more than a predefined percentage (as given by k) of the cell's total value. Note also that the $k\%$ differs for $n = 1$ and $n = 2$ which makes sense as it allows for the independency between the two rules in finding cells having high risk of disclosure. If we by contrast consider the two dominance rules with the same k as in the current example, i.e. $(1,80)$ and $(2,80)$, we are introducing an unnecessary and redundant level of dependency between the rules. Consider that according to the rule $(1, 80)$, the cell is found to be sensitive (or, unsafe), that is the largest contributor contributes more than 80% of the cell's total. In this case there is no need to check whether the cell is sensitive according to the $(2, 80)$ rule as it will necessarily be so, which follows from the first rule.

If, by contrast, we assume that the cell is safe according to the first rule, it will be safe or unsafe according to the second rule depending on whether the contribution of the two major contributors surpass 80%. Moreover, if we come to the conclusion that the cell is safe according to the both rules mentioned above, it can still be possible for an intruder to identify a close upper estimate of the contribution of the remaining competitors. If we assume that the intruder is the one having the largest contribution, and that her contribution is large enough but smaller than 80% of the total, she can subtract her own value from the cell's total and still get a close upper estimate of the contributions of the remaining contributors, despite the fact that the cell was marked as safe. Therefore, setting the same k for the two dominance rules based on $n = 1$ and $n = 2$ seems redundant.

2.3. Given that the response variable is non-negative, explain why a minimum frequency rule "at least 3" isn't needed if we have checked with the dominance rule (2, *k*) or the *p*%-rule (1,*p*)?

Minimum frequency rules are generally employed if it is considered that preventing exact disclosure is sufficient. The minimum frequency rule postulates that the cells based on at least as many respondents as some minimum frequency n are considered safe. In most cases, $n = 3$.

If some variables are considered highly confidential, the minimum frequency rules may not suffice and other rules, such as dominance rules and p%-rules are recommended. Moreover, the minimum frequency rule with $n = 3$ is redundant if the dominance rule $(2, k)$ and p%-rule $(1, k)$ are applied given that the response variable is non-negative. This is because both the dominance rule $(2, k)$ and p%-rule $(1, k)$ account for additional information about the contribution of the second largest contributor (on par with the largest contributor) when assessing the cell for the disclosure risk. These rules imply that the number of contributors is at least 3.

When the response variable can take on negative values, a minimum frequency rule is often preferred to the dominance and p%-rules (Hundepool *et al.*, 2010).

Problem 3

3.1. Describe the difference between deterministic and stochastic protection methods, and nonperturbative and perturbative protection methods.

Stochastic and deterministic methods. Stochastic methods employ a probability mechanism or a random number-generating mechanism to protect the data. That's why stochastic methods always produce different outcomes whenever the methods are applied to the data. A general recommendation is to set a seed for the random number generator in order be able to produce replicable results when needed. Deterministic methods, by contrast, follow a deterministic algorithm, that is, they always produce the same results (with the obvious caveat that they are applied to the same data and the same set of parameters). Recoding and local suppression, for example, are deterministic methods, while swapping and PRAM are stochastic methods.

Non-perturbative and perturbative methods. Deterministic and stochastic methods can in turn be classified as non-perturbative and perturbative. Non-perturbative methods work by reducing the details in the data either by generalization or by suppression of certain values. Perturbative methods, as the name suggests, work by perturbing the values in a dataset, and not by suppressing them. The goal of the perturbation is to induce uncertainty around the true values. Recoding and local suppression, for example, are non-perturbative deterministic methods, while PRAM and swapping are perturbative stochastic methods. Other perturbative methods are micro-aggregation, addition of additional noise, etc. (Templ *et al.*, 2020)

3.2. Where would you say cell suppression fits in these classifications?

Cell suppression occurs when certain sensitive information about a particular individual or individuals is suppressed by, e.g., replacing the corresponding cell values in a dataset by missing values. Thus, I would say that cell suppression is a deterministic and non-perturbative method.

3.3 Give some examples other than cell suppression

One such method, for example, is global recoding. It is a non-perturbative deterministic method which can be applied to both categorical and continuous variables. The idea is to combine several categorical variables into one category so that the category is less informative than the individual categories alone. A common example is to categorize ages into groups. Continuous variables may also be combined into groups (intervals). In this case it is said that they are categorized, i.e., they are treated as discrete (categorical) variables. A common example is categorization of income. It follows that the larger (the broader) the groups, the less informative they are. And at the same time, the larger the groups, the greater is the loss of information.

A popular example of perturbative deterministic methods is micro-aggregation. Typically, this method is applied to continuous variables. The method works by splitting the records into groups; for each group its arithmetic mean (or some other more robust measure) is calculated. These values are then used to replace the individual values in each group.

Another popular example of perturbative protection methods is addition of noise to the original values. This method is used for microdata and applied to continuous variables. The aim of adding noise to the data is protect it from exact matching with external files (*Templ et al.*, 2020).

Problem 4

PRAM has been mentioned during lectures and by its inventors suggested as a method of protection. Explain shortly how it is done and in what circumstances it is applicable

PRAM is a stochastic perturbative method, also known as a post-randomization method, which is recommended mainly for categorical variables.

The main idea of the method is to change the original values of the categorical variables in a dataset into some other (possible) categories according to the probabilities established *a priori*. Each value of a categorial variable receives a predefined probability to be changed into some other category. To be able to apply the method, a transition matrix holding the probabilities for each category or level of the categorical variable should be established.

For example, if a categorical variable with, say three levels (A_1, A_2, A_3) is to be perturbed by the PRAM method, the transition matrix will be given by

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix},$$

where p denotes the probabilities of A_1, A_2, A_3 of being re-classified as another level of the categorical variable. The matrix is read row-wise. The first row of the matrix, for example, represents the probabilities of A_1 to be re-classified as either A_2or A_3. The second and the third rows represent the probabilities of A_2 to be re-classified as either A_1 or A_3, and of A_3 as either A_1or A_2 respectively. Note, that all the row-wise probabilities must sum up to 1. For instance, if the probability of A_1 to remain A_1 is, say 0.7 ($p_{11} = 0.7$), and the probability of A_1 to be re-classified as A_2 is 0.2 $(p_{12} = 0.2)$, then the probability that A_1 will be re-classified as A_3 should be equal to 0.1 $(p_{13} = 0.1)$.

PRAM is a stochastic method. Since it is applied independently and randomly to each observation, each run of the method will produce different results unless the seed from each run is explicitly stored.

PRAM is considered a flexible protection method as it allows to model any desired effects by specifying and modifying the probabilities of a transition matrix. Thus, if there is no need to change a level of the categorical variable into some other level due, for example, to its commonality, it is fully legitimate to set the transition probability of the level to 1, and the probabilities of all the other levels to 0. For example, in the matrix above, the probability $p_{11} = 1$ would mean that the category A_1 remains A_1 with probability 1, that is it stays unchanged. This necessitates that the probabilities for A_1 be changed to A_2 and A_3 must be set to 0, i.e., $p_{12} = p_{13} = 0$ (Benschop *et al.*, 2019).

Reference list

Benschop T., Machingauta C., Welch M., The World Bank, 2019; https://readthedocs.org/projects/sdcpractice/downloads/pdf/latest/

Carlson M., Some contributions to statistical disclosure control, Department of Statistics, Stockholm University, 2002

Hundepool A. *et al*., Handbook on statistical disclosure control, Version 1.2, ESSNet SDC, 2010

Lambert D., Measures of disclosure risk and harm, Journal of Official Statistics, Vol. 9, No. 2, 1993, pp. 313-331

OECD Glossary of Statistical Terms, https://stats.oecd.org/glossary/detail.asp?ID=6932

Templ M., Meindl B, Kowarik A. Introduction to Statistical Disclosure Control, 2020; https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf

‘