Statistical Methods - Part 3

Philip Dahlqvist-Sjöberg

5^{th} of June 2020

1 Disclosure

Statistical disclosure (unraveling), is defined as when information about a individual or an organization, i.e. subject, is unwillingly shared to other subject trough new data. The data is most commonly published in either frequency tables, or magnitude tables. The information at risk, can be categorized in to two types; identity disclosure and attribute disclosure (Hundepool et al. 2010, p. 9). These problems are most common for National Statistical Institutes (NSI), due to the types of studies and trust they inflict on society.

Identity disclosure is defined as when the identity of a subject can be associated with published data, when the subject's identity is believed to be anonymous. The difference with attribute disclosure, is when the identity of a subject's participation in the study is not necessarily confidential, but specific attributes of that subject can be associated with the published data, linked to the subject.

The point of statistics is to aggregate and summarize relationships of data. Hence, the published data in it's self, might not be at risk of disclosure, however, information could be inferred with high confidence from the published data, thus occurs inferential disclosure (OECD 2005).

For example, the study and hence published data, might investigate the effect from number of children on amount of received welfare of a family. If there is a high correlation between number of children and amount of welfare, information about number of children for a specific family, can infer with high confidence the amount of welfare that specific family receives.

The information about amount of welfare is sensitive information, and thus the trust between the subjects and NSI can be violated if inferential disclosure occurs. However, OECD (2005) argues that most NSI's do not care as much about inferential disclosure, due to their purpose of enabling society with data to infer relationships with.

Unique sets of key attributes, is when there is only one subject attaining a curtain set of key attributes. This is a scenario of disclosure risk, and can be represented in two categories of identifications; spontaneous recognition and recognition via matching/linkage (Hundepool et al. 2010, p. 37).

The identity of a subject within a study, can easily be identified if a unique sets of key attributes of that subject is published. Hypothetically, if I posses knowledge of identifying attributes of a friend, whom without my knowledge has anonymously participated in a national study about child abuse, I could identify this friend if the data table published, have a subject with a unique set of key attributes, that matches my friend's identifier attributes. I could then identify him/her in the table (Hundepool et al. 2010, p. 37). This is a case of unique set of key attributes, that inflicts identity disclosure.

However, there might also be additional attributes. Maybe I was not aware that this friend had been abused as a child etcetera, hence providing me new information about this friend, violating the condition of anonymity and disclosing attributes with a unique match, instead of a full set of unique key attributes. Hence, it is not a sufficient nor a necessary condition of uniqueness for attribute disclosure, merely a close link of attributes can provide me confidence of disclosure. An example of this is inferential disclosure described above.

2 Cell risk of disclosure

a)
The dominance rule, (n, k)-rule, is defined as,

$$Unsafe - cell := x_1 + \cdots + x_n > \frac{k}{100}X = \frac{x_1 + \cdots + x_n}{X} > k\%,$$

where $x_1 \ldots x_n$ is ordered x-values from the total X. Hence with the first rule, $(1, 60)$, we use only the largest contributor in each cell,

$$Unsafe - cell := \frac{x_{largest}}{X} > 60\%.$$

Looking in Table 1, the unsafe cells have been marked with dark red. These cells, have a single contributor, that represents more than 60% of the entire cell value. Hence, this cell is at risk of being manipulated by subtraction or coalition, in order to harm the anonymity of the subjects in the cell. For the second rule, $(2, 90)$, the dominance rule uses the two largest values,

Table 1: Dominance rule $(1, 60)$ and $(2, 90)$, from 2a). The dark red cells have been marked as unsafe cells based on both rules. The light red, has been marked additionally as unsafe cells from only the second rule, $(2, 90)$. The blue cell has been marked as unsafe, but from the zero sum rule. The purple cell has been marked as unsafe too, but from the subtraction of attributes of the North marginal.

	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

$$Unsafe - cell := \frac{x_{largest} + x_{second}}{X} > 90\%.$$

Looking in Table 1, the unsafe cells have been marked with dark and light red, where the light red is additional cells compared to the first rule. This result have used the information about the two largest contributors in each cell, to identify unsafe cells. Here, the two largest contributors, cover more than 90% of the entire value for the marked cell.

With knowledge of the two largest contributors, these cells are sensitive to subtraction, hence, it is necessary to take protective actions against disclosure of information.

b)
The p%-rule, (m-1,p), measures how contribution from $m - 1$ subjects can estimate the contribution from the largest and left out contributor. The rule is defined as,

$$Unsafe - cell := \frac{X - (x_1 + \cdots + x_m)}{X} < p\%.$$

For the case of $(1, 11)$, the formula hence look like,

$$Unsafe - cell := \frac{X - (x_{largest} + x_{second})}{X} < 11\%,$$

where the second largest, can estimate the largest. Looking in Table 2, the unsafe cells have been marked with red. This result shows which cells, that have the second and largest contributor, representing almost the entire value of the cell. These two contributors have more than 90% of the contribution, hence the largest contributor can be identified by the second largest contributor by subtraction.

Table 2: P%-rule $(1, 11)$, from 2b). The red cells have been marked as unsafe cells. The blue cell has been marked as unsafe, but from the zero sum rule. The purple cell has been marked as unsafe too, but from the subtraction of attributes of the North marginal.

	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

c)
For dominance rules, with different number of contributors, n, but with equal percentage, k, each rule with increasing number of contributors, will always contain the information of the previous rule, as well as additional information.

In the case of $(1, 80)$ and $(2, 80)$, each cell, where *one* contributor attains 80% of the cell value, *two* contributors will also always attain 80% of the cell value. Hence, the first rule is redundant when you have the second rule.

d)
The minimum frequency rule uses the knowledge of each cell's contributors frequency, to determine if a cell is unsafe of disclosure (Hundepool et al. 2010, p. 119). The level of unsafe frequency is set usually too, at least three.

The dominance rule checks the contribution of n subjects. Hence, for a dominance rule of $(2, k)$, the total number of subjects is at least 2, and for any $k < 100\%$, this rule will mark all cells with fewer than three subject by default, fulfilling the minimum frequency rule of at least three.

From the definition described in b), the p%-rule $(m-1, p)$ uses $m-1$ subject to estimate the left out subjects. For $m - 1 = 1$, $m = 2$ subjects. Hence, if a cell have less than three subjects, and $p > 0\%$, the cell will be disclosed by default.

Both the dominance and p%-rule are types of concentration rules (Hundepool et al. 2010, p. 121), hence both the dominance $(2, k)$ rule, and the p% $(1, p)$ rule, by default fulfill the minimum frequency rule of at least three.

3 Disclosure protection

When it comes to statistical disclosure, there are different distinctions of protection methods; deterministic and stochastic (probabilistic) methods are two types of categories (Templ et al. 2020). Both methods have pros and cons when trying to prevent disclosing information in the published data.

The deterministic method uses a standard way of protecting the information e.g., always round the value with the basis of b. Hence deterministic method is less effective, because if the attacker of the information attains the standard protection and b, in this case, the attacker can reverse the protection very easily.

Further, the deterministic method will most likely inflict problems when, for example, aggregating information in a table, when the sum of a row, do not add up to it's marginal. If instead the marginal is calculated after rounding, the aggregated value can differ a lot from the true value, hence providing additional unnecessary loss of information.

The stochastic method, uses a probabilistic approach to the protection. For the rounding example, it could have a probability p of round up to the closest value with base b, and down similarly with probability $1-p$. Hence, this method is more effective, but may inflict bias unless p is set so, expected value for each subject is it's original value.

Protection methods are also categorized in two other approaches; perturbative and non-perturbative protection methods (Hundepool et al. 2010, p. 54). A perturbative method, alter the data before publication, in order to create confusion which preserves the confidentiality of the data. For a successful perturbative method, the data should be altered in such a way, that the attained statistics are significantly non-different between the original and altered data. In contrast, a non-perturbative method do not alter the data before publication. However, it produce partial masking of the data, in such a way that it protects from disclosure.

Cell suppression is a commonly used protection method. It fits in the categories; non-perturbative and deterministic methods (Templ et al. 2020). This method uses a determined level of safety, for example number of subjects in each cell, in order to make the assessment for protection or not. If protection is needed, the information within this cell, and sometimes other effected cells, are removed without altering all data.

Other methods in the same categories as suppression, is Global recoding and Top and bottom coding (Templ et al. 2020).

Global recoding, combines multiple categories of a categorical variable or intervals for continuous variables, in order to "hide" the unsafe cells in a larger group of subjects. For example, different age groups can be created, in order to avoid unique sets of key attributes in a table.

Top and bottom coding, is similar to Global recoding, but works for ordered variables. This method only recode the values of the lowest and highest values, which implies that it works best when the distribution of subjects, are centered over the ordered scale. This method protects the outlier subjects, which often have the most sensitive information, and are the easiest to identify.

4 PRAM

Hout & Elamir (2006) article on Post Randomization Method (PRAM), describes a perturbative protection method from disclosure for categorical variables. The method uses a fixed probabilistic approach to reclassify some categorical variables in the data. This method is very effective in protecting the information in the data, since even if an intruder matches subjects, the identification is only identified with a certain probability.

The probabilities, defined by a transition matrix $\boldsymbol{P}$, is published together with the data, in order to allow consistency between the estimations before and after PRAM. This is achieved if $\boldsymbol{P}$ is set, so that the expectation of the PRAM values are equal to original values.

The strength in this method, lies in it's ability to protect the data from disclosure, wile simultaneously reducing the information loss as much as possible. The user of the data, say a researcher, can apply invariant PRAM, which is calculating the expected value of the data, which yield unbiased estimates (Hout & Elamir 2006, p. 727), so statistical analyses are very close to estimates on the original data. However, when applying statistical disclosure, information loss in inevitable (Hout & Elamir 2006, p. 712), but this method reduces the loss significantly.

Hout & Elamir (2006) continues to discuss situations when PRAM is reasonable to apply in practise. In many cases, tables published from NSI's are of interest for the general public, where precise details are not the objective. However some e.g., researchers might be interested in specific details in the data. When using other protection methods these details may disappear, but the PRAM method can be a solution to keep the details requested from the researcher.

The uses of remote data access is ever so increasing, not only in industries implementing cloud services etcetera, but also for NSI's. Hout & Elamir (2006) describes a data quarry process, where the method is applied on the requested data. This way, the data is not only secure form hackers, but can avoid a subset weakness for the method. If a subset of the perturbed data is preferred, when PRAM has already been applied to the full data, additional estimations are required to attain the invariant PRAM. If quarries are requested from the remote server, PRAM method can be applied directly on that subset, with it's unique $\boldsymbol{P}$ matrix. Hence, the additional estimations are avoided. All unnecessary steps in the process are preferably removed, because of how complex several dimensions in the data can become.

Lastly, they discuss how the PRAM method requires low correlation between the categorical variables in the data. Introducing PRAM on highly correlated data, will inflict inconsistencies, which might lead to increased risk of disclosure.

References

Hout, A. V. D. & Elamir, E. A. (2006), 'Statistical disclosure control using post randomisation: Variants and measures for disclosure risk', *Journal of Official Statistics* **22**(4), 711–731.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G. & Wolf, P.-P. D. (2010), *Handbook on Statistical Disclosure Control*, ESSnet.

OECD (2005), 'Inferential disclosure'.
URL: *https://stats.oecd.org/glossary/detail.asp?ID=6932*

Templ, M., Meindl, B. & Kowarik, A. (2020), 'Anonymization methods'.
URL: *https://sdcpractice.readthedocs.io/en/latest/anon_methods.html*