Stockholm University

Master thesis

Department of Statistics

Masteruppsats, Statistiska institutionen

Comparison of different sensitivity rules in primary cell suppression for tabular data

Qun Wang

Masteruppsats 30 högskolepoäng, vt 2013

Supervisor: Dan Hedlin

Table of Contents

Abstract

Statistical disclosure control (SDC) is a set of methods that are used to reduce the risk of disclosing information on individuals, businesses or other organizations. There is a wide range of different techniques for SDC according to different kinds of data, while here the focus is on sensitivity rules about how to define whether a cell in the tabular data has the risk of disclosing information. The current sensitivity rules include the dominance rule, the p% rule and the p-q rule. However, we have discovered some problem with the p% rule and Professor Johan Bring who is the founder of the statistical consultancy "Statisticon" has come up with a new method and named it the "Statisticon rule".

There are many discussions in the previous literature regarding which rule is more conservative, i.e. which rule classifies more cells to be sensitive. However we believe that no rule is more conservative than the other, and simply by changing the value of the parameters, each rule can be adjusted equally conservative. Instead of focusing on distinguish which rule is more conservative, we attempt to detect when the different rules classify a cell to be sensitive rather than how often they classify cells to be sensitive in this thesis.

Pairwise comparisons based on simulation were made and one of the most important results indicates that when the number of sensitive cells are set up to be the same, the dominance rule and the p% rule have a lot in common when it comes to whether to classifying a cell to be sensitive, but the Statisticon rule and the p% rule differ a lot from each other. The p% rule tends to classify a cell to be sensitive when it has only one observation with extremely big value, and the Statisticon rule tends to classify a cell to be sensitive when it has two observations with big values.

When there are strong reasons to believe that it is more dangerous when the true value of the biggest contributor to a cell is closer to the upper boundary of the estimated value that the second largest contributor can assume, the p% rule is a better choice. Otherwise, the Statisticon rule is more general since it also takes into consideration that the interval of the largest contributor that the second largest contributor could calculate and is therefore more preferred than the Dominance rule and the p% rule.

Key words: statistical disclosure control, tabular data, cell suppression, sensitivity rules, simulation

Acknowledgements

I would like to give my special gratitude and deepest respects to both of my wonderful supervisors – Professor Dan Hedlin from the Statistic Institute of Stockholm University and Professor Johan Bring from Statisticon, I appreciate the opportunity they gave me to get to know this brand new area in statistics for me, also I am thankful to their guidance and support through out the whole process.

I am indebted to Doctor Feng Li from the Statistic Institute of Stockholm University for his valuable help in R.

I am grateful to Professor Michael Carlsson from the Statistic Institute of Stockholm University and Doctor Ingegerd Jansson from Statistics Sweden for their precious suggestions and share of experiences.

I also want to thank my dear friend Tania Arria, for her accompany and encouragement which made the last semester very pleasant.

1. Introduction

This chapter is used to discuss the general background and the main goal of conducting this research. Furthermore, a brief outline on the orgnization of this thesis is given.

1.1Background

National Statistical Institutions (NSIs) are expected to provide our society with trustworthy and detailed statistical outputs, however these statistical outputs may sometimes lead to disclosure of some sensitive and confidential information, which might cause harm on both individual level and organization or group level. For the individuals this is more likely related to privacy whereas for organizations or companies, the connection is more connected with issues such as commercial secrets and obviously this is not the intention of publishing the data.

In order to protect the confidentiality of survey respondents – not only because of legal and ethical mandates, but because public trust and perceptions of that trust are important contributors to data quality and response rates (Doyle *et al*, 2001), NSIs have come up with a set of different methods to protect the confidentiality of the information provided by the respondent. A very good expression quoted from Statistics Sweden about the aim of statistical disclosure control is "to show the forest instead of the tree".

1.1.1 What is disclosure risk?

A disclosure occurs when a person or organization recognizes or learns something that they did not know already about another person or organization, via released data according to Duncan *et al* (2001). Such person or organization that intends to know something about others is referred to as *intruder*.

There are three types of disclosure risk: identity disclosure, attribute disclosure and inferential disclosure. Identity disclosure occurs with the association of a respondent's identity with a disseminated data record containing confidential information (Duncan *et al* 2001). In other words, an *intruder* could identify a specific individual or organization according to the statistics published.

Attribute disclosure occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent (Duncan *et al* 2001). This is to say an *intruder* could determine the value of some survey variable for an identified individual or organization using the statistical output (Skinner, 2009).

Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of a home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject (OECD Glossary of Statistical terms, 2013).

1.1.2 What is Statistical Disclosure Control (SDC)?

In order to prevent such disclosure from happening, statistical institutes are responsible to utilize some approaches to reduce the disclosure risk and further to protect the confidentiality.

Statistical disclosure control (SDC) techniques can be defined as the set of methods that are used to reduce the risk of disclosing information on individuals, businesses or other organizations (Hundepool *et al*, 2010). There is a wide range of different techniques for SDC according to different kinds of data. More detailed information about the approaches is going to be discussed in the method/theory chapter. The main goal of SDC is to find a balance point where the disclosure risk is reduced within a tolerable level and meanwhile the information provided by data is as much as possible.

1.1.3 How does SDC influence the data?

Using SDC will apparently disturb the information of the data, which means the utility is therefore decreased. *Figure 1* is a map of disclosure risk (R) and utility of the data (U);

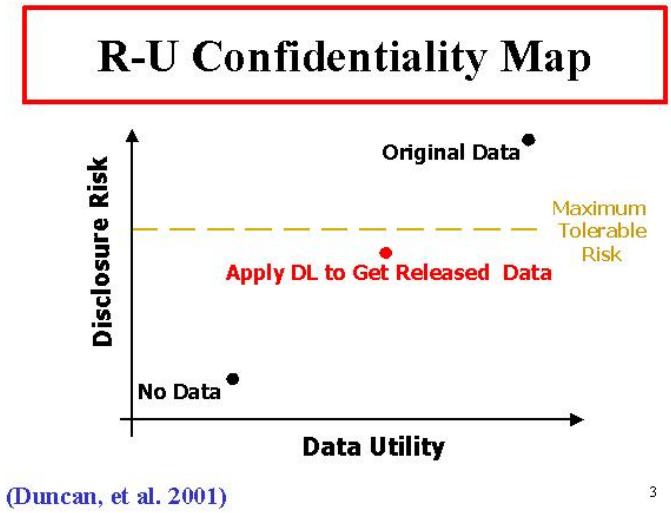


Figure 1 R-U Confidentiality map

The original data has the largest utility but also the highest disclosure risk. On the contrary, if no data were published then there would be no risk of disclosure; however, this disobeys the responsibility of statistical institutions. The released data should therefore be, with the help of SDC approaches, located somewhere under the maximum tolerable risk line and have the utility as large as possible. In a word, the goal is to minimize disclosure risk while maintaining the analytic utility of the data.

1.2 Research question and aim

The main focus of this thesis is on one of the most popular SDC approaches for tabular data —cell suppression. The research question and the aim of this thesis is to compare the different sensitivity rules for cell suppression at the primary level and thereafter find out in what conditions are these sensitivity rules suitable to be applied.

1.3 Organization of the thesis

This thesis begins with introducing the background and purpose of the research question. Thereafter in the theory chapter, different sensitivity rules of the approach “cell suppression” in SDC are presented. Additionally a new rule is introduced and formulated. Afterwards a

chapter about comparison between the different rules based on simulation comes and the results are discussed. The thesis ends with a discussion chapter, where all the results and conclusions are summarized and discussed.

2. Theory

2.1 Micro-data and Macro-data

Generally speaking, the data that NSIs publish are comprised of two kinds--micro-data and macro-data. Micro-data Files are data sets containing for each respondent the score on a number of variables (Hundepool *et al*, 2010). It is easily understood that such data have high possibility of both identity disclosure and attribute disclosure since it contains details for every single respondent.

The disclosure rate could be managed by SDC or by restricting access of the data. There are very strict restrictions when distributing private data in Sweden. It is illegal to publish any individual's information and the regulations are very rigid even for research purposes (Public Access to Information and Secrecy Act (Offentlighets och sekretesslagen), 2009). Hence micro-data is not the focus and is not going to be discussed in this thesis.

Macro-data on the other hand, which could also lead to disclosure and especially group disclosure, is the main focus of this research. Magnitude table and frequency table are two major components when displaying macro-data.

2.2 An example on tabular data disclosure

It might be non-intuitional that disclosure happens for such aggregate forms as well. This is possibly due to e.g. the object in the cell is unique. However, uniqueness is not a necessary condition for disclosure. *Table 1* shows a simple artificial example illustrates the situation:

Table 1 Level of income at University S (artificial data)

Departement	Low Income	High Income	Total
A	8	4	12
B	10	1	11
C	1	0	1
D	5	0	5
Total	24	5	29

There is only 1 person (unique) working at department C with a low income. Therefore, the income level is disclosed as long as it is known that she/he works at department C. Similarly, it is easily deduced that all people working at department D have low income too, despite the object in the cell low income is 5 (not unique). This is called group disclosure. When taking a look at department B, there are 10 people with low income and 1 with high income, therefore, the one who has the high income would know that she/he is the only one in this department who has high income. The disclosure happened within the group could sometimes be more harmful than disclosure from the outside.

2.3 SDC approaches

There are different options available for SDC approaches when a cell in a table is identified as sensitive. Generally speaking, the SDC approaches could be divided into two categories, namely perturbative methods and non-perturbative methods.

Perturbative methods allow for the release of the entire micro data set, although perturbed values rather than exact values are given, for example adding noise to micro-data and then the table is formed from the perturbed micro-data. (Domingo-Ferrer and Torra, 2001)

Non-perturbative methods do not rely on distortion of the original data but on partial suppressions or reductions of detail, for example global coding, which means a table can be redesigned by combining sensitive cells with other cells. (Whillenborg and De Waal, 2001) Alternatively the value of sensitive cells can be suppressed or modified so that the table comprises less detailed information (Hundepool *et al*, 2010), this is called cell suppression.

In other words, perturbative methods changes the original data and the data released are not all true values. On the contrary, the data released using non-perturbative methods are true values but with less detailed information.

2.4 Cell suppression

Although there are many different approaches concerning SDC, the focus of this thesis is on cell suppression. Cell suppression does not conduct the protection by changing the information of the data but by subtracting detailed information of the data. Therefore, it is a non-perturbative method.

Cell suppression is a popular and efficient SDC approach. However, it is important to recognize that it has some drawbacks. First of all, there might be some unnecessary information loss when applying cell suppression, though information loss is a general drawback for all SDC approaches. Secondly, some follow-up problems might occur after cell suppression. A simple artificial example is given to clarify this problem, *Table 2* is about the level of subsistence allowance for different age group in city A：

Table 2 Level of subsistence allowance in city A (artificial data)

Age Groups	100 %	50 %	25 %	Total
20-30	1	0	0	1
30-40	12	3	5	20
Above 40	24	12	4	40

There is no column total published with this table. The object for age group 20-30 with subsistence allowance level at 100% is only 1 and should therefore be primarily suppressed, the row total for that age group should also be secondarily suppressed. However, the next month, some other resources publish the total number of people in city A having 100% subsistence allowance, and one could combine these two tables together to disclose the suppressed cell. Therefore, when applying cell suppression, there is always risk when some supplementary information is published.

There are two steps in the cell suppression procedure. The first step is called “primary suppression” by using different criteria, namely sensitivity rules, to determine which cells are primary unsafe. If a table presents marginal totals, which is usually the case, then there is an

additive relationship between the cells of the table. Therefore, it is not enough to consider disclosure risk on the level of individual cells only, but additional cells must be suppressed as well. This second step is called "secondary suppression" or "complementary suppression" (Hundepool *et al*, 2010). To be more precise, this thesis goes deeply into comparison of different sensitivity rules in "primary suppression" in particular.

Before penetrating different sensitivity rules, let $x_1 \geq x_2 \geq \ldots \geq x_N$ be the ordered contribution in a cell by respondents 1,2,…,N respectively and let $X = \sum_{i=1}^{N} x_i$. A discussion about the most well-known sensitivity rules concerning "primary suppression" is given as following:

2.4.1 The dominance rule

The dominance rule is also called the (n, k) rule. The cell is primarily considered unsafe if the total of the n largest contributions exceeds k% of the total cell value X

That is the cell is unsafe according to this rule if and only if:

$$x_1 + x_2 + \ldots + x_n > k/100 * X \quad (1)$$

2.4.2 The p% rule

The p%-rule is based on the following inequality:

$$X - x_j - x_i < p/100 * x_i \quad (2)$$

The worst case of this inequality is to suppose x_i as the largest respondent and x_j as the second largest respondent, where the largest respondent is estimated by the second largest. The assumption for the "worst case" is therefore: x_2 knows it is the second largest contributor and is trying to estimate the largest contributor x_1. As long as this condition is safe, the cell is safe for all other conditions. For the p% rule, a cell is considered sensitive if the cell total minus the 2 largest contributions is smaller than a certain percentage of x_1, namely,

$$X - x_2 - x_1 < p/100 * x_1 \quad (3)$$

If the cell satisfies the above inequality, the cell is sensitive from the viewpoints of all respondents. The assumption for both dominance rule and p% rule is that there is no prior knowledge about the contribution for respondent x_i. More detailed information could be found in (Loeve, 2001).

2.4.3 Comparison of the dominance rule and the p% rule

It is believed in many literatures that in general, the p% rule is preferred to the dominance rule because the dominance rule is more conservative than the p% rule, i.e. it provides more suppressions of the cell than the p% rule. For example in "Handbook on statistical disclosure control" (Hundepool *et al*, 2010) and "Statistical disclosure control in tabular data" (Castrol, 2009), the dominance rule with n=2 is compared with p% rule, since when n=2 the dominance rule looks very similar to formulation of the p% rule.

The (2,k) rule classifies a cell as sensitive if

$$x_1 + x_2 > k/100 * X \Leftrightarrow X - x_2 - x_1 < (1 - k/100) * X. \quad (4)$$

Comparing (3) and (4), it is seen that in both cases a cell is sensitive if $(X - x_2) - x_1$, i.e. the difference between the estimation of x_1 made by second respondent and the true value of x_1, is less than a certain percentage of either the first respondent value x_1 in (3) or the cell value X in (4). Indeed, for p and k such that $k = 100 * \frac{100}{100+p}$, every non-sensitive cell for the rule (2,k) is also a non-sensitive cell for the p% rule; but the reverse implication does not hold.

Above are the arguments about why the dominance rule is not as good as the p% rule. However, none of the articles give the motivation about why to limit the analysis to $k = 100 * \frac{100}{100+p}$. When this equation doesn't hold, these two rules are not comparable any more. Each rule can be made equally restrictive by changing the value of parameters and hence it is very hard to tell which rule is better. A more detailed comparison based on simulation that support this argumentation is going to be discussed in the next chapter.

2.4.4 The p-q rule

The P-q rule is an extension of the p%-rule, which is also called prior posterior rule. As stated above, there is no prior knowledge about the contribution for respondent x_i. In the p-q rule, on the contrary, all respondents in a cell are assumed to know the values of the other contributions to that cell within at most q percent (Loeve, 2001). The cell is sensitive according to the p-q rule if

$$p * x_1 - q * \sum_{t=3}^{N} x_t \geq 0 \quad (5)$$

This rule is not commonly applied since it is impractical in reality. It is very difficult to indicate how much knowledge all respondents know about the other contributions, i.e. to evaluate q.

2.4.5 Drawback with the p% rule

Suppose a cell is the sum of three respondents A, B and C. Assume B has the share 40 and the total sum is 100. These are all pieces of information known to B and B could calculate the interval of the largest company (either A, C or B itself) is (40, 60). There are many different combinations of the values for A and C. For example the values of A, B and C could be (59, 40, 1) or (41, 40, 19)

According to the p% rule in inequality (3), the cell is identified as sensitive when the values of A, B and C are (59, 40, 1) and safe when (41, 40, 19), which seems that when A (or C) is closer to the upper bound (which is 60 in this example) of the interval calculated above the cell is considered sensitive, on the contrary, when A (or C) lies near by the lower bound of the interval (which is 40 in this example) the cell is considered safe. This is problematic because the share of the B has stayed unchanged with value 40 and information that B remains the same. There is no additional knowledge about the other respondents in case of (59, 40, 1) than

in (41, 40, 19). It can be argued that it is unreasonable that the former is sensitive but the latter is safe. The following *Figure 2* illustrates this problem.

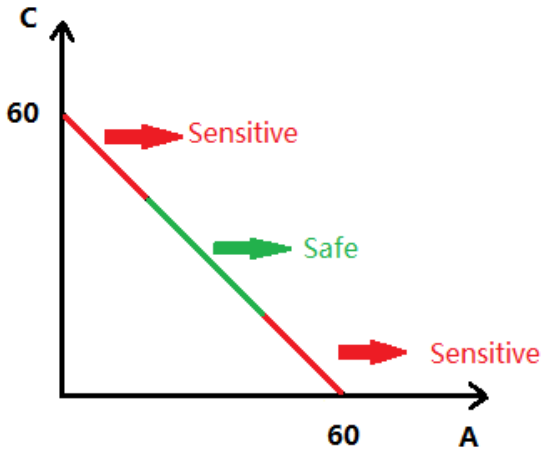


Figure 2 whether a cell is sensitive or safe depending on the distribution of A and C. Assuming X = 100 and B = 40

When either A or C is closer to the upper bound (which is 60 in this example), the cell is sensitive and otherwise the cell is safe. However, 59 and 41 have the same distance to the limits of the interval. This is unreasonable because from B's point of view, it should not make any difference whether A (or C) is 41 or 59, the disclosure would happen despite the share of the A (or C). What makes a difference is the size of the interval where A (or C) is located.

It is assumed in the book "Handbook on statistical disclosure control" (Hundepool *et al*, 2010) that when there are no coalitions of respondents, the closest upper estimate of the largest contributor can be obtained by the second largest contributor, by subtracting her own contribution x_2 from the aggregate total X, i.e. $\hat{x}_1 = X - x_2$. If the difference between the estimated $\hat{x}_1$ and the true value x_1 is too small, i.e. $X - x_2 - x_1 < p/100 * x_1$, then the cell is considered sensitive.

However, this assumption is questionable, take the above example again: the second largest contributor B = 40 and the aggregate X = 100, then B knows the largest contributor lies between 40 and 60. If no other information obtained, a smart guess would be around 50, there is no supporting evidence for B to estimate $\hat{x}_1 = X - x_2 = 60$ because the probability that the largest contributor is 60 is much lower than that for 50. From this aspect, the p% rule is not theoretically solid since it is based on the assumption that $\hat{x}_1 = X - x_2$.

2.4.6 The Statisticon rule

Concerning the problem of the p% rule, Professor Johan Bring has come up with a new sensitivity rule, which is based on the information that the intruder knows. Professor Bring is

the founder of the statistical consultancy "Statisticon", therefore he named the new rule the "Statisticon rule".

The principle of the Statisticon rule is that the decision if a cell is safe or sensitive should only be based on the information from the intruders' perspective. If the intruder respondent j is trying to estimate another respondent i, say x_j is estimating x_i , based on the information that intruder knows, it is possible to calculate the interval I where respondent i may lie in, i.e. $I_{x_i} = x_{i\,max} - x_{i\,min}$. If this interval is very narrow, there is risk for disclosure. A cell is classified sensitive when:

$$I_{x_i} < \text{s}/100 * \text{X} \tag{6}$$

The "worst case" here is again when supposing x_i is the largest respondent and x_j is the second largest respondent, where the largest respondent is estimated by the second largest. The assumption for the "worst case" is therefore: x_2 knows it is the second largest contributor and is trying to estimate the largest contributor x_1. As long as this condition is safe, the cell is safe for all other conditions. For the Statisticon rule, a cell is considered sensitive if the Interval for x_1 is smaller than a certain percentage of X, namely,

$$I_{x_1} < \text{s}/100 * \text{X} \tag{7}$$

Now the question is how to calculate $I_{x_1} = x_{1\,max} - x_{1\,min}$. We could easily estimate that $x_{1\,max}$ = X - x_2, and for $x_{1\,min}$ is a bit more complicated:

$$x_{1\,min} = \begin{cases} \text{X} - (\text{n}-1) * x_2 & when\ \text{X} - (\text{n}-1)\,x_2 > x_2,\ \text{namely}\ x_2 < \text{X}/\text{n} \\ x_2 & when\ x_2 \geq \text{X}/\text{n} \end{cases} \tag{8}$$

Let us review the example in 2.4.5 again: suppose a cell is the sum of three respondents A, B and C. Assume B has the share 40 and the total sum is 100. Now B is in the position of x_2 and trying to estimate x_1. Then $x_{1\,max}$ = $\text{X} - x_2 = 100 - 40 = 60$ and since 40 > 100/3 ➔ $x_{1\,min} = x_2 = 40$ and the interval $I_{x_1} = x_{1\,max} - x_{1\,min} = 60 - 40 = 20$. If we set s = 25, we get 20 < 25/100 *100, and the cell is sensitive; if we set s = 15, we get 20 > 15/100 * 100, and the cell is safe. Different from the P% rule, the Statisticon rule classifies a cell to be sensitive only on the interval of x_1 and different value of s. It does not matter whether the combination of the values for A and C is (59, 40, 1) or (41, 40, 19), because the information that B knows has not changed at all.

3 Comparisons between sensitivity rules based on simulation

In order to compare the different sensitivity rules in a deeper and more intuitive way, simulations are introduced. In this chapter, the simulation setup and the conclusions draw from analyzing the result of simulation are going to be presented.

3.1 Simulation Setup

Two different types of distributions are used for simulation, namely the exponential distribution and normal distribution.

The reason why to simulate from exponential distribution is due to its characteristic: suppose x~exp ($\lambda = 1$), the frequency decreases as the value of x increases, the histogram is presented in *Figure 3*:

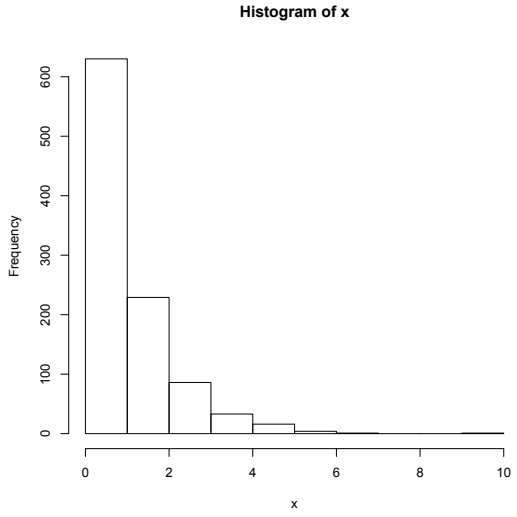


Figure 3 Histogram of x, where x is simulated for 1000 times and follows exponential distributed with $\lambda = 1$

The feature of exponential distribution represents a plausible scenario of the cases for disclosure – when there is one or two extreme values the possibility of disclosure is high. Normal distribution is also included in simulation as an alternative distribution because it is the most common distribution and many data follow normal distribution.

1) **Fundamental construction**

   Create an N*n matrix **X**, where M is the total amount of simulation and let N = 1000. Each row of **X** is referred as a "dataset". Each dataset contains n observations ($x_1, x_2, \dots, x_n$) which follow one of the above-mentioned distributions, where n ranges from 3 to 20, is the number of observations in each dataset of the matrix **X**. The aggregate $X = \sum_{i=1}^{N} x_i$ for each dataset is referred as a "cell". As mentioned in the theory chapter, $x_1 \geq x_2 \geq \dots \geq x_n$ are the ordered contribution of the dataset by respondents 1,2, …, n respectively.

Here is an example of the first ten datasets in matrix **X** when n=4 and x is generated from exponential distribution with $\lambda = 1$. Note here the observations (x_1, x_2, x_3, x_4) are not ordered.

*Example 1 First ten datasets (the total amount of datasets is M = 1000) in matrix **X** when n=4 and x follows exponential distribution with* $\lambda = 1$

	[,1]	[,2]	[,3]	[,4]
[1,]	1.87	0.22	0.91	1.63
[2,]	0.40	1.05	0.68	4.42
[3,]	0.15	1.31	1.78	0.24
[4,]	1.73	0.22	1.93	0.60
[5,]	0.09	2.49	0.07	0.69
[6,]	0.67	0.13	1.09	1.05
[7,]	1.07	0.62	0.77	0.25
[8,]	1.51	1.57	0.26	1.16
[9,]	1.31	1.08	0.07	0.24
[10,]	0.16	2.33	2.12	0.20

2) Definition of the rules

Define the three sensitivity rules, namely the dominance rule, the p% rule and the Statisticon rule according to the formulas mentioned in the theory chapter. For convenience's sake, these three rules are denoted as D, P and S respectively.

3) Build sensitivity matrix for each rule

Use the defined rules above to construct a matrix for each rule that conveys 1000 rows and m columns, where m represents the parameter in each rule, namely k in the D, p in the P and s in the S, and they range from 1 to 100. The value of each elements in the matrices are either 1 or 0, where 1 represents the dataset is identified as sensitive by the corresponding sensitivity rule, and 0 otherwise. Thus three matrices, **Dmatrix**, **Pmatrix** and **Smatrix** are formulated, all of them are 1000 by 100 matrices with values of 1s and 0s. Here the number of observations n in each dataset is fixed to be 4.

Here is an example of the first ten rows and 14 columns in matrix **Pmatrix** when n=4 and x is generated from exponential distribution with $\lambda = 1$.

Example 2 First ten rows and 14 columns in matrix ***Pmatrix*** *when n=4 and x follows exponential distribution with* $\lambda = 1$

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	1	1	1	1	1	1	1	1
[6,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0

For example, the 5th row means the 5th dataset, where the first 6 elements are 0 and the rest are 1, indicating that the dataset is sensitive if $p \geq 7$. Hence, this matrix tells us that for which values of p the dataset is classified as sensitive and the column sum shows how many datasets are classified as sensitive for each value of p.

Draw a plot for the three rules simultaneously, where the x-axis represents different values of m and y-axis is each column sum, which means the number of sensitive datasets identified by different rules.

4) Pairwise comparison when n is fixed

In order to make pairwise comparison between the rules, we should first make sure that these three rules are comparable. The number of sensitive datasets y represents the degree of confidentiality of the data. When the total number of sensitive datasets classified by different rules are the same, that is, (sensitive datasets / total datasets) | D = (sensitive datasets / total datasets) | P = (sensitive datasets / total datasets) | S, we could assume the three rules are equally restrictive as each other. This could be achieved by changing the value of the parameter.

We should first determine the number of observations n in each dataset and the total number of sensitive datasets y. In this case, we choose n = 4 and y = 200 for all three rules. For example when y = 200 out of the total simulated datasets 1000 means 20% of the datasets are sensitive for all three rules. Only when the number of observations and the confidentiality of the data are fixed, we can start the pairwise comparison. Other values can be chosen as needed.

Thereafter, the pairwise comparisons between the rules could be made. Calculate the value of parameters that make the total number of sensitive datasets equal 200 for each rule, we get k = 89, p = 18, and s = 27. Even though the number of sensitive datasets are equal for each rule, which datasets are the sensitive ones are not necessarily the same for all three rules. The idea is to pick out all the datasets that are classified as sensitive by one rule but as safe by another and analyze the characteristics of such datasets.

Take the comparison between P and S as an example, construct a matrix **A** containing all the dataset that are calculated to be "1" according to P but at the same time calculated as "0" according to S. Similarly, construct matrix **B** containing all the dataset that are calculated to be "1" for S but at the same time calculated as "0" for P. Afterwards, plot all the 4 observations in each dataset in **A** and **B** separately and inspect if there is any pattern shown. (Note here that the elements should be presented in percentage instead of the real value for a better comparison.) Briefly put, matrix **A** contains all the datasets that are identified to be sensitive by P but safe by S and matrix **B** contains all the datasets that are identified to be sensitive by S but safe by P.

*Example 3 First ten rows in matrix **A** where all the datasets are identified to be sensitive by P but safe by S and **B** where all the datasets are identified to be sensitive by S but safe by P, when n=4 and x follows exponential distribution with* $\lambda = 1$

```
      [,1] [,2] [,3]  [,4]
 [1,] 0.09 2.49 0.07  0.69
 [2,] 4.49 0.04 1.24  0.74
 [3,] 1.70 2.94 0.17  0.32
 [4,] 1.59 0.28 0.06  4.15
 [5,] 0.00 0.82 1.73  0.07
 [6,] 0.19 0.05 6.94  1.45
 [7,] 0.13 0.01 0.89  0.33
 [8,] 0.02 1.19 0.44  0.05
 [9,] 0.07 1.77 0.90  0.18
[10,] 0.27 0.01 0.03  1.68
```

*Matrix **A***

```
      [,1] [,2] [,3] [,4]
 [1,] 0.15 1.31 1.78 0.24
 [2,] 1.73 0.22 1.93 0.60
 [3,] 1.31 1.08 0.07 0.24
 [4,] 0.67 0.11 1.56 1.76
 [5,] 0.80 0.03 1.41 1.59
 [6,] 0.18 1.10 0.36 0.96
 [7,] 0.80 0.44 0.80 0.03
 [8,] 0.85 1.84 0.02 2.11
 [9,] 0.28 1.60 0.09 0.20
[10,] 2.57 0.99 2.95 0.00
```

*Matrix **B***

In order to compare the variability and analyze the characteristic of matrix **A** and **B**, we can calculate the standard deviation for each dataset and then calculate the average standard variation for all the datasets in **A** and **B** respectively. Additionally, calculate the difference between the largest and the second largest element in each dataset for **A** and **B** respectively. The same procedure is repeated for comparing P and D as well as S and D.

5) Contour plot of the number of sensitive datasets when n and m are unfixed

In the previous stages, the value of n and y are fixed, naturally, the corresponding value of m for each rule can be calculated and are also fixed values for all of the three rules. As a matter of fact, m, n and y compose a three-dimensional variation, the value for n, y and m can change simultaneously, and therefore a contour plot can give a good overview of this variation.

Construct three matrices **D.AmountSensitive**, **P.AmountSensitive**, and **S.AmountSensitive** for each rule correspondingly, where the row is the number of elements in each dataset, n (ranges from 3 to 20 or more if needed), the column is the

value of m (ranges from 1 to 100), so there are in total 18 rows and 100 columns. The elements in the matrices are the number of sensitive datasets y classified by each rule in the corresponding m and n value. Thus, these three new build matrices are 18 by 100 dimensioned. The contour plot should be drawn over these three matrices in order to give a general sight about how the total number of sensitive datasets changes as the number of element in each dataset n and the value of the parameter m changes for each rule.

Example4 First 10 rows and 12 columns in matrix ***P.AmountSensitive*** *when x follows exponential distribution with* $\lambda = 1$

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	8	0	0	0	0	0	0	0	0	0	0	0
[2,]	13	0	0	0	0	0	0	0	0	0	0	0
[3,]	19	0	0	0	0	0	0	0	0	0	0	0
[4,]	23	0	0	0	0	0	0	0	0	0	0	0
[5,]	26	0	0	0	0	0	0	0	0	0	0	0
[6,]	29	0	0	0	0	0	0	0	0	0	0	0
[7,]	30	0	0	0	0	0	0	0	0	0	0	0
[8,]	34	0	0	0	0	0	0	0	0	0	0	0
[9,]	35	0	1	0	0	0	0	0	0	0	0	0
[10,]	39	2	1	0	0	0	0	0	0	0	0	0

In this example, the columns represent n, which ranges from 3 to 20, that is to say, [,1] represents n=3, [,2] represents n=4 and etc. Let us take the element in the first row and first column 8 as an example, it means when n=3 and p=1, there are 8 datasets classified as sensitive according to the P.

6) Extension

The total number of simulation N, the value of the number of elements in each dataset n, and the number of sensitive datasets y can all be modified into values according to the specific need. Same procedure could be done for the normal distribution or even some other distributions.

3.2 Result and discussion

The results obtained from the previous section are presented in this chapter. Note here we assume the number of elements n in each dataset is 4 and x follows exponential distribution with λ=1. The results for normal distribution and other values of n could be referred to Appendices.

3.2.1 Sensitivity plot

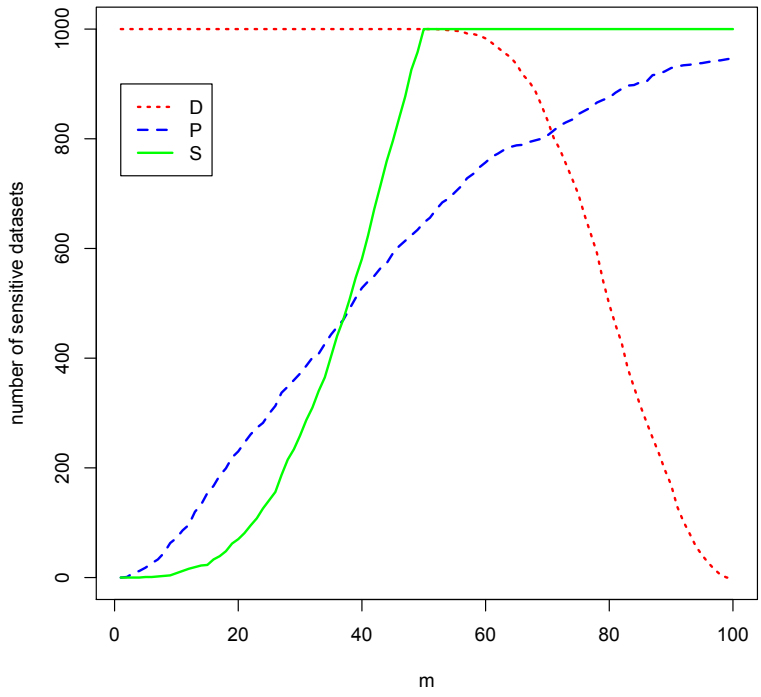


Figure 4 Sensitivity plot: number of sensitive datasets plot of D, P and S as the value of m ranges from 0 to 100, when n=4 and x follows exponential distribution with λ=1

It is worth mentioning again that here n = 4 and the value could be modified if needed. The x-axis represents different values of m and y-axis is the number of sensitive datasets identified by different rules.

The number of sensitive datasets decreases as the m increases for D, which is reasonable since it is consistent with the definition of the Dominance rule – the cell is identified as sensitive if the two largest shares of the cell are bigger than a certain value of m. If m is too small, almost all cells have the two largest shares bigger than m.

For P and S, the situation is just the reverse. The value of the number sensitive datasets increases as the value of m increase. But the pattern for increasing for this two rules are not identical, and when m reaching a certain value, all of the datasets are classified as sensitive for S but that never happens to P even when m = 100.

This plot only shows the pattern of the number of the datasets that are identified as sensitive by each rule, however, there is no information about the characteristics of the sensitive datasets for each rule, for example, whether the datasets that are sensitive for D are the same ones as for P or S.

3.2.2 Contour plot of the number of sensitive datasets of D, P and S

Courant *et al* (1996) stated, "A contour of a function of two variables is a curve along which the function has a constant value." The contour plot is a plot illustrated with contour lines. The lines represent the number of datasets classified as sensitive by each rule, the x-axis is the value of the parameter m (ranges from 1 to 100), and y-axis is the number of elements n in each dataset, ranges from 3 to 20. This contour plot shown in *Figure 5* illustrates explicitly the trend of how the number of sensitive datasets **y** varies with the number of observations in a dataset **n** and the value of the parameter **m** (k for D, p for P and s for S).

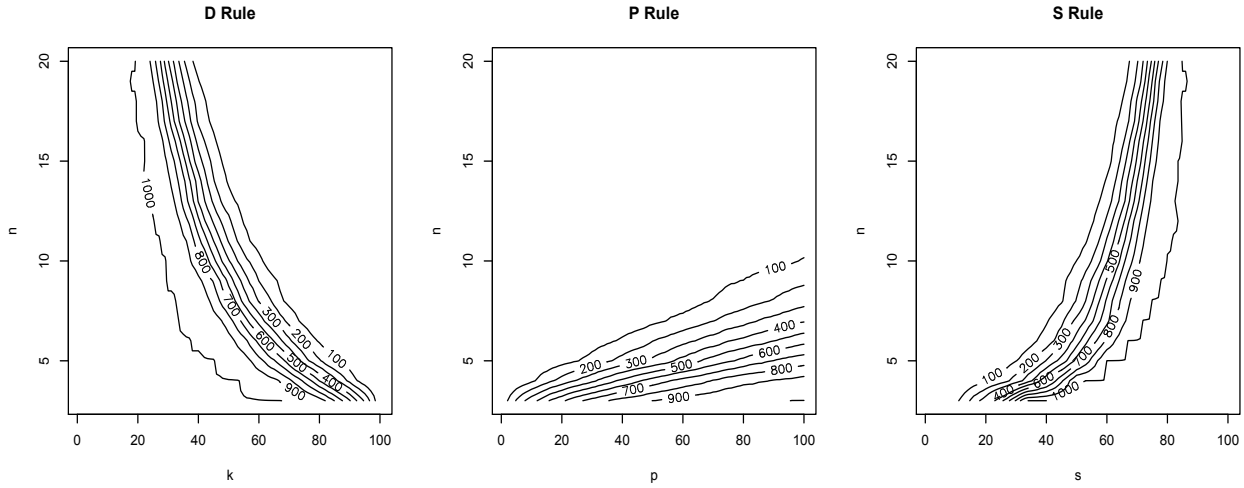


Figure 5 Contour plot: the number of sensitive datasets of D, P and S, when number of observations n in a dataset ranges from 3 to 20, and the value of parameter m ranges from 0 to 100 and x follows exponential distribution with λ=1

Some conclusions could be drawn:

i. For all three rules, the smaller n is, the greater y is in general. For P and S, the bigger m gets, the greater y is, but it is opposite for D. When n is greater than 10, none of the datasets in P is sensitive but this characteristic does not appear for D and S.

ii. The sensitivity lines for D and S are symmetric around the y-axis indicating that the trend of y varies with n and m are very similar between D and S. Even though the trend for D and S are similar, which datasets are the sensitive ones are not necessarily the same for them. Therefore, the conclusion that D and S are equivalent could not be achieved.

3.2.3 Pairwise comparison of P and S

As we have discussed in the previous sub-chapter, in order to make the pairwise comparison between the rules, the number of sensitive datasets y should be fixed to be the same value. In the next step, the number of sensitive datasets y = 200, and the different values of parameter m that make the total number of sensitive datasets equals 200 are calculated for each rule, where k = 89, p = 18 and s = 27. Thereafter, the pairwise comparisons between the rules could be made. Even though the number of sensitive datasets is equivalent for each rule, which datasets are the sensitive ones are not necessarily the same for all. To begin with, the comparison between P and S is carried out (note here that the number of sensitive datasets for P and S are 199 and 187 respectively, but not exactly 200. This is due to the fact that k, p and s are rounded to the nearest unit), *Table 3* shows the distribution of the sensitive datasets classified by P and S respectively:

Table 3 The distribution of the sensitive datasets classified by P and S

	S → Sensitive	S → Safe	Total
P→ sensitive	75	124	199
P → safe	112	689	801
Total	187	813	1000

Even though the total number of sensitive datasets is limited to be the same for both P and S, in most conditions, these two rules are not equivalent: there are only 75 datasets that classified to be sensitive by both rules. 112 datasets are classified as sensitive by S but not by P (classified to group A), and 124 datasets are classified as sensitive by P but not by S (classified to group B).

How does this table describe the extent of agreement between these two rules? According to Viera and Garrett (2005), the kappa coefficient is the most commonly used statistic for measuring agreement between two or more observers. A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. However, there is no clear cut-off point for kappa coefficient, Viera and Garrett (2005) suggest the following table, which may help to "visualize" the interpretation of the kappa coefficient in *Table 4*:

Table 4 Interpretation of the Kappa coefficient

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	0.2	0.4	0.6	0,8	1.0

The Kappa coefficient calculated between P and S is 0.24, which is close to a slight agreement, indicating that the extent of agreement between these two rules is low. In this case, it is very interesting to take a deep look at these datasets and detect how they differ in group A and B.

We start with giving a simple typical case in group A and B respectively by taking the median of the observations x_1, x_2, x_3 and x_4 in every dataset (note here that the observations should

be presented in percentage instead of the real value for a better comparison), a typical case for each group is provided in *Table 5*:

Table 5 Typical cases for group A (sensitive by P but safe by S) and B (safe by P but sensitive by S)

	Typical case
A (sensitive by P but safe by S)	$x_1 = 0.69, x_2 = 0.23, x_3 = 0.06, x_4 = 0.02$
B (safe by P but sensitive by S)	$x_1 = 0.45, x_2 = 0.38, x_3 = 0.12, x_4 = 0.05$

It could be seen that in group A, namely the datasets that are classified as sensitive by P but safe by S, the value of x_1 is much greater than all the other elements in the same dataset, meanwhile in group B, where the datasets are classified as safe by P but sensitive by S, the value of x_1 and x_2 are both quite large compared with the rest elements in the same dataset.

Plot these two typical cases so that we can get a more intuitive grasp in *Figure 6*:

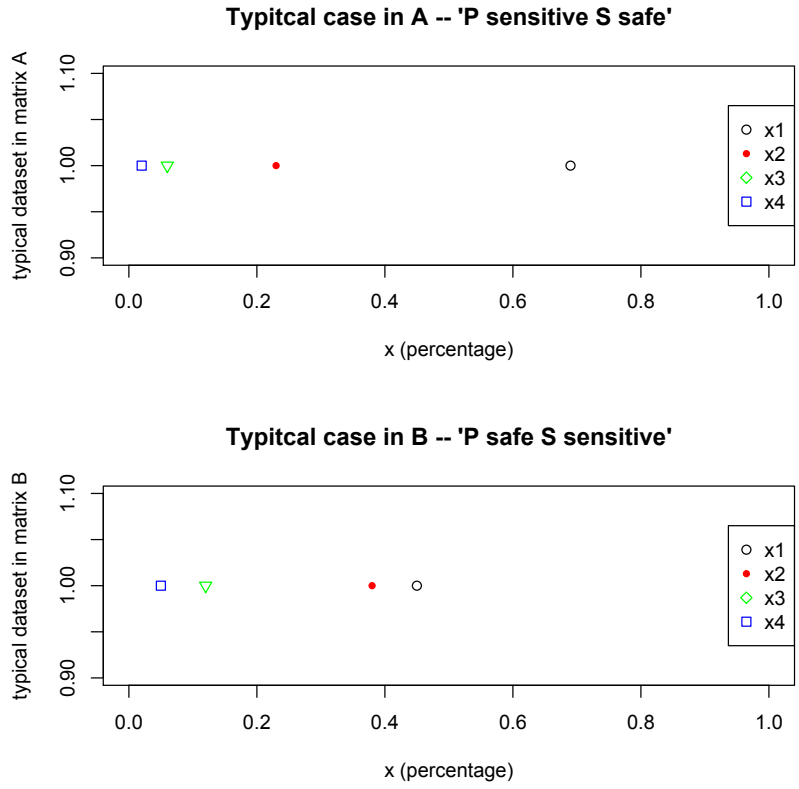


Figure 6 Plot of x_1, x_2, x_3 and x_4 of the typical case for A and B, when x is generated from exponential distribution with λ=1,number of observations n = 4, and the number of sensitive datasets y = 200

The same conclusion could be drawn by inspecting *Figure 6*, in group A, x_1 is far away from the other elements, however in group B, x_1 and x_2 are quite close to each other and far from

the rest of the elements. We now want to detect if all the datasets in group A and B show the same pattern, this could be realized by drawing the same plot for all the datasets for group A and B respectively. A more general grasp of these two groups is illustrated in *Figure 7*:

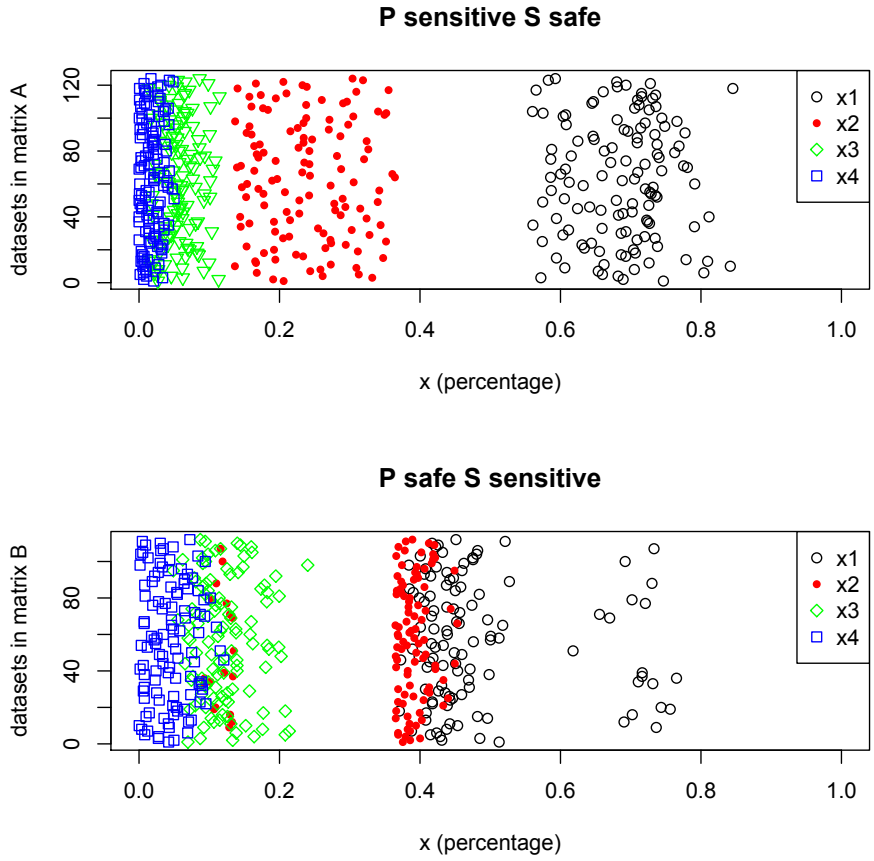


Figure 7 Plot of x_1, x_2, x_3 and x_4 for all datasets in A and B, when x is generated from exponential distribution with λ=1,number of observations n = 4, and the number of sensitive datasets y = 200

It is quite obvious that *Figure 7* shows the same pattern as in *Figure 6*: in group A, x_1s are extremely large (greater than 50%), the percentage for x_2s, x_3s and x_4s are close to each other; on the contrary, in group B, the share of x_1s and x_2s blend together, despite there are some extremely large outliers for x_1. If the above two plots only give an intuitionistic overview, then the next step proves this conjecture in a more concrete way. Calculate the average standard deviation of datasets the and calculate the difference between x_1s and x_2s for group A and B respectively which are shown in *Table 6*:

Table 6 The average of the standard deviation and the average of the difference between x_1 and x_2 for all datasets A and B

	Mean (sd)	Mean (Difference between x_1s and x_2s)
A (sensitive by P but safe by S)	1.36	1.98
B (safe by P but sensitive by S)	0.83	0.53

Table 6 gives the same conclusion: The average difference between the corresponding x_1s and x_2s in A is much bigger than that in B, and the average standard deviation in A is greater than that in B. To be more exact, P tend to classify a dataset to be sensitive when it has one observation with extremely big value, and S tend to classify a dataset to be sensitive when it has two observations with big values.

3.2.4 Pairwise comparison of D and S

The same procedure repeated for comparing D and S as well as P and D. And the results are as following:

Table 7 The distribution of the sensitive datasets for D and S

	S → Sensitive	S → Safe	Total
D→ sensitive	78	118	196
D → safe	109	695	804
Total	187	813	1000

The distribution of the sensitive datasets for D and S is quite similar to P and S in *Table 3*: there are only 78 datasets that are classified to be sensitive by both rules. 118 datasets are classified as sensitive by D but not by S, and 109 datasets are classified as sensitive by S but not by D. The Kappa coefficient calculated between D and S is 0.27, which is also close to a slight agreement, indicating that the extent of agreement between these two rules is low.

Repeat the same procedure when comparing D and S, construct two matrices R and U, where R contains all the datasets that are classified as sensitive by D but safe by S, and U contains all the datasets that are classified as sensitive by S but safe by D. Afterwards, plot these datasets for R and U respectively in *Figure 8*:

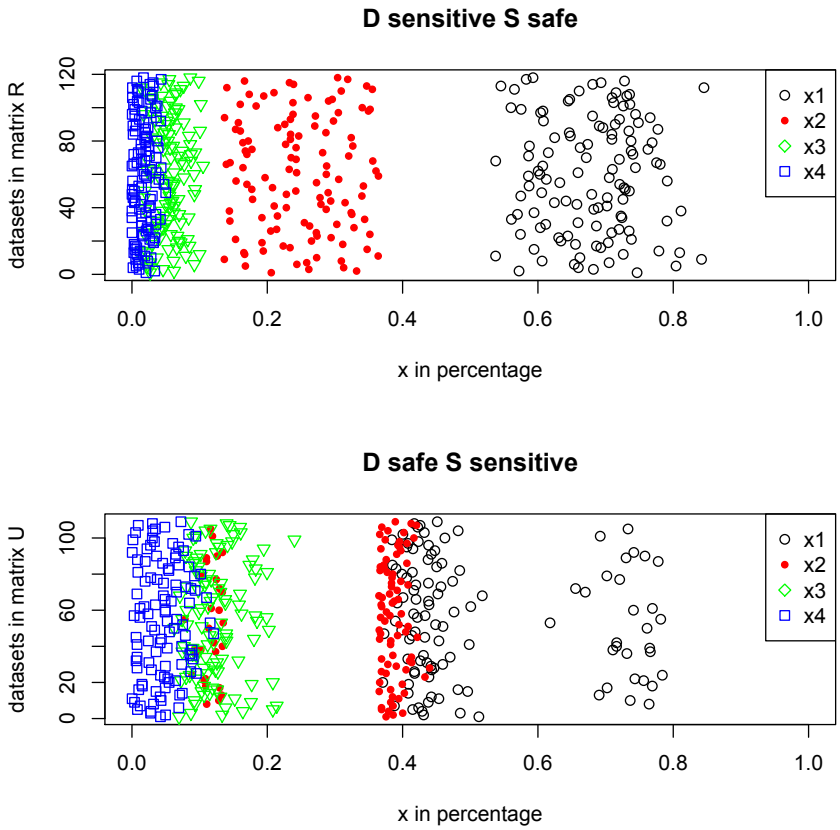


Figure 8 Plot of x_1, x_2, x_3 and x_4 for all datasets in R and U, when x is generated from exponential distribution with λ=1,number of observations n = 4, and the number of sensitive datasets y = 200

Figure 8 looks quite similar as *Figure 7* as they almost have the same pattern, we could hence draw a same conclusion: D tends to classify a dataset to be sensitive when it has one observation with extremely big value, and S tend to classify a dataset to be sensitive when it has two observations with big values.

3.2.5 Pairwise comparison of P and D

We repeat the above procedure once again to compare P and D, and the results are as following:

Table 8 The distribution of the sensitive datasets classified by P and D

	D→ Sensitive	D → Safe	Total
P → sensitive	177	22	199
P → safe	19	782	801
Total	196	804	1000

There are 177 datasets that are classified to be sensitive by both rules. Only 22 datasets are classified as sensitive by P but not by D and only 19 datasets are classified as sensitive by D but not by P. Unlike the comparison between P and S, these two rules are equivalent at most

of the times, in total, there are only 41 datasets that are classified differently. The kappa coefficient is calculated to be 0.81, indicating a substantial agreement between these two rules.

Repeat the same procedure when comparing P and S, construct two matrices G and H, where G contains all the datasets that are classified as sensitive by D but safe by P, and H contains all the datasets that are classified as sensitive by P but safe by D. Afterwards, plot these datasets for G and H respectively in *Figure 9*:

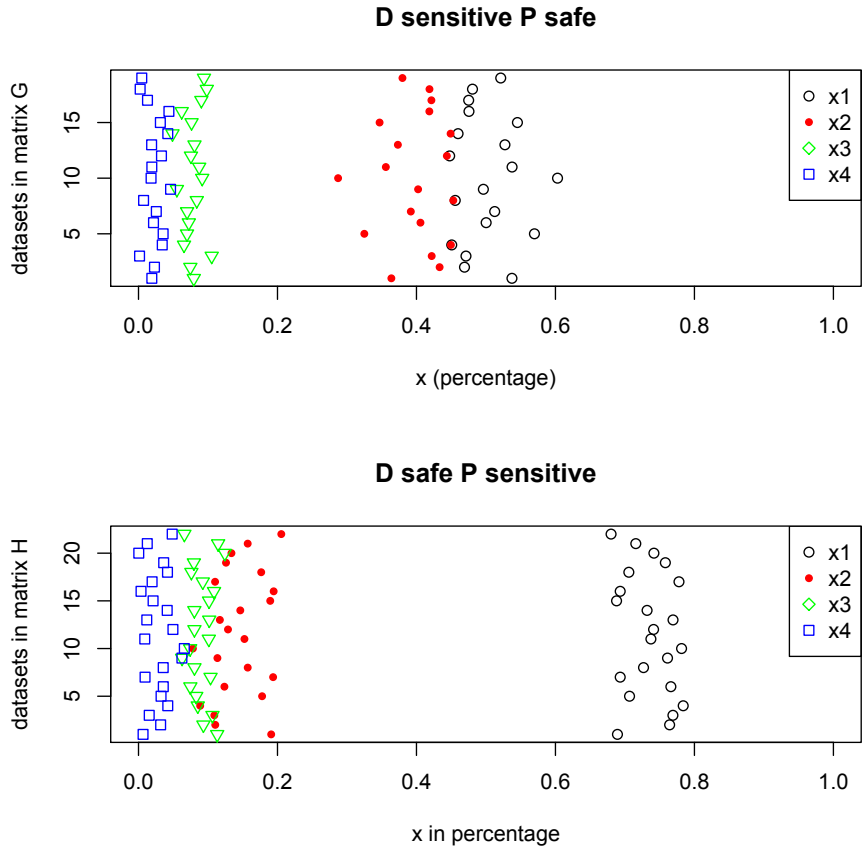


Figure 9 Plot of x_1, x_2, x_3 and x_4 for all datasets in G and H, when x is generated from exponential distribution with λ=1,number of observations n = 4, and the number of sensitive datasets y = 200

Even though there are only a few datasets that are classified differently by P and D, the pattern is quite obvious. P tend to classify a dataset to be sensitive when it has one observation with extremely big value, and D tend to classify a dataset to be sensitive when it has two observations with big values.

4. Discussion

There are many discussions in the previous literature regarding which rule is more conservative, i.e. which rule classifies more cells to be sensitive. However, as we have discussed in the former sections: no rule is more conservative than the other, and simply by changing the value of the parameters from 0 to 100, each rule can be adjusted equally conservative so that it can classify the cells as sensitive to a certain amount.

Instead of focusing on distinguish which rule is more conservative, we put more emphasis on how the data looks like and what distribution it has. That is why we attempt to detect when the different rules classify a cell to be sensitive rather than how often they classify cells to be sensitive in this thesis.

The most important conclusion is that the p% rule tends to classify a dataset to be sensitive when it has one observation with extremely big value, and S tends to classify a dataset to be sensitive when it has two observations with big values. Recall the typical cases when comparing the p% rule and the Statisticon rule in *Table 5*:

Table 5 Typical cases for group A (sensitive by P but safe by S) and B (safe by P but sensitive by S)

	Typical case
A (sensitive by P but safe by S)	$x_1 = 0.69, x_2 = 0.23, x_3 = 0.06, x_4 = 0.02$
B (safe by P but sensitive by S)	$x_1 = 0.45, x_2 = 0.38, x_3 = 0.12, x_4 = 0.05$

Dataset (0.69, 0.23, 0.06 and 0.02) is classified as sensitive by the p% rule, recall the formulation of P, a cell is considered as sensitive if X – x_2 – x_1 < p/100 * x_1, that is 1 – 0.69 – 0.23 = 0.08 < 18/100 * 0.69 = 0.12, so this dataset is sensitive according to the P% rule, this is because 0.69 is very close to the upper boundary of x_1, where $\hat{x}_1$ = X – x_2 = 0.77. This corresponds to the problem with the p% rule that we have discussed in chapter 2.4.5. If the value of x_1 in this dataset is not 0.69 but 0.59, while the dataset sum remains 1 (so $x_3 + x_4$ would gain 0.1), the dataset is no more sensitive by the p% rule: 1 – 0.59 – 0.23 = 0.18 $\not<$ 18/100 * 0.59 = 0.11. However, the information that x_2 knows has not changed at all, why does it matter if x_1 is 0.69 or 0.59? One possible explanation is that some times, it might be more dangerous when true value of x_1 is closer to the upper boundary of the estimated $\hat{x}_1$ that x_2 can assume. When does this situation occur should be determined by the more experienced statisticians or experts in NSIs and the specific area where the data comes from. The Dominance rule and the P% rule are quite similar in this respect.

Nevertheless, the Statisticon rule is more general, as it classifies this dataset as safe because the interval of x_1 calculated by x_2 is (0.23, 0.77) which is quite wide. The Statisticon rule does not only take the situation when true value of x_1 is closer to the upper boundary of the estimated $\hat{x}_1$ that x_2 can assume into consideration, but also the interval of x_1 that x_2 could calculate. If there is no strong reason to believe that it is more dangerous when true value

of x_1 is closer to the upper boundary of the estimated $\hat{x}_1$ that x_2 can assume, the Statisticon rule is more preferred than the Dominance rule and the P% rule.

References

Castro J. (2009), *Statistical disclosure control in tabular data (Report DR 2009-11)*, Universitat Politècnica de Catalunya

Courant, R., H. Robbins, and I. Stewart, (1996) *What Is Mathematics?An Elementary Approach to Ideas and Methods,* New York: Oxford University Press, pp. 344

Domingo-Ferrer, J. and Torra, V. (2001), *Disclosure Control Mehods and Information Loss for Microdata*. In Doyle, P., J.I. Lane, J.J.M. Theeuwes and L.V. Zayatz (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier, pp. 91-108

Doyle, P., J.I. Lane, J.J.M. Theeuwes and L.V. Zayatz (eds.) (2001) *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies*, Amsterdam: Elsevier, pp. 1

Duncan, G.T., S.E. Fienberg, R. Krishnan, R. Padman and S.F. Roehrig (2001) *Disclosure limitation methods and information loss for tabular data.* In Doyle, P., J.I. Lane, J.J.M. Theeuwes and L.V. Zayatz (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier, pp. 135-166

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, E. S. Nordholt, G. Seri and P. P. De Wolf (2010). *Handbook on statistical disclosure control.* ESSNet SDC, Eurostat

Loeve, J. A. (2001), *Notes on sensitivity measures and protection levels,* Project number: TMO-102966, Statistics Netherlands.

Office for national statistics UK (2012), *GSS/GSR Disclosure Control Policy for Tables Produced from Surveys.* Available at: *http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html*

Skinner C.J. (2009), *Statistical disclosure control for survey data (Working paper M09/03),* Southampton Statistical Sciences Research Institute, Southampton, UK

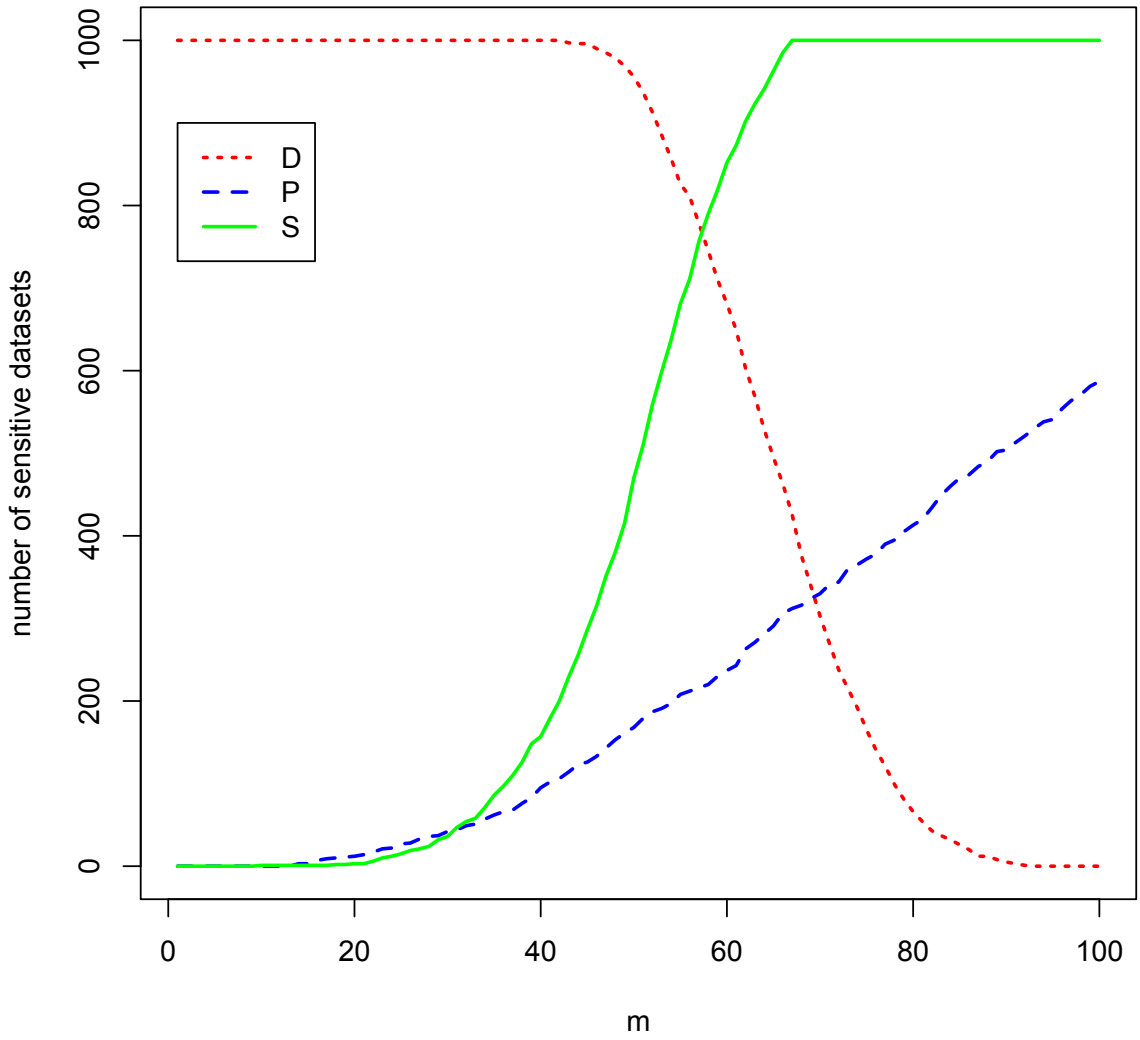
Sveriges Riksdag (2009), *Public Access to Information and Secrecy Act (Offentlighets och sekretesslagen)*, chapter 24 §8, available at: *http://www.notisum.se/rnp/sls/lag/20090400.htm*

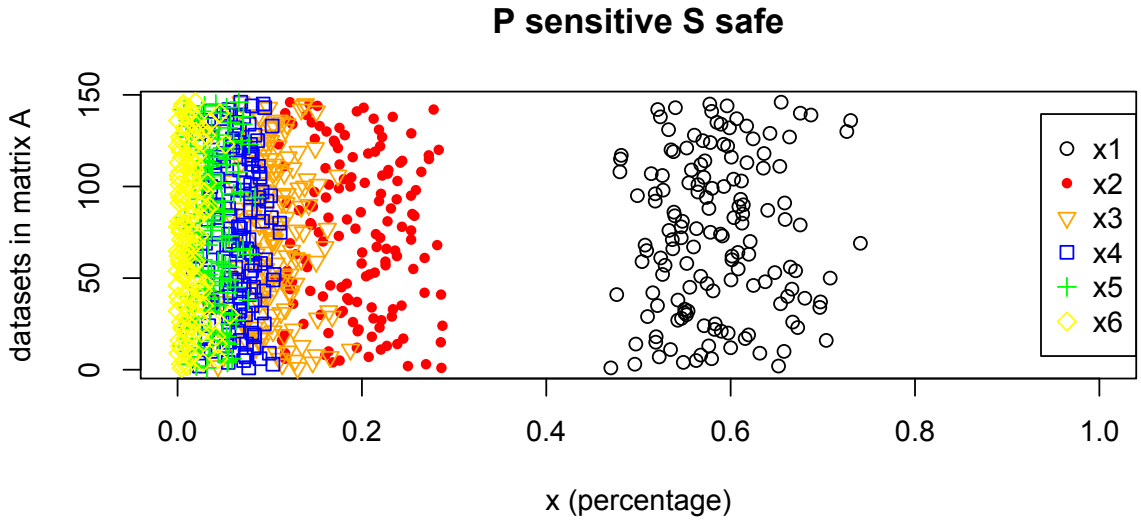
The Organization for Economic Co-operation and Development (OECD) official website *(*Retrieved May 3 2013*), the OECD Glossary of Statistical Terms*, available at: *http://stats.oecd.org/glossary/detail.asp?ID=6932*

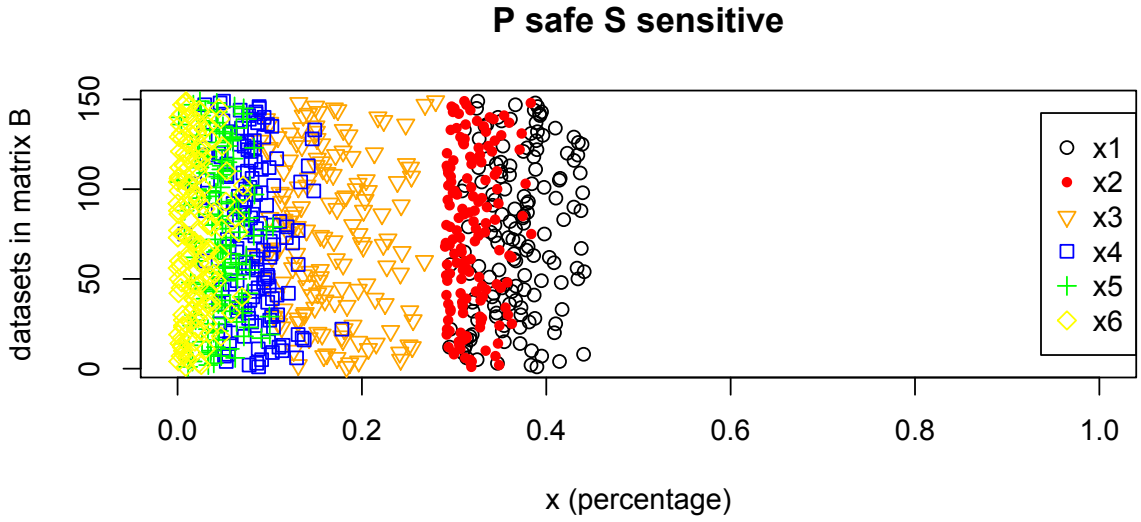
Viera, A.J. and J.M. Garrett (2005), *Understanding inter observer agreement: the kappa statistic,* Source: Robert Wood Johnson Clinical Scholars Program, University of North Carolina, USA

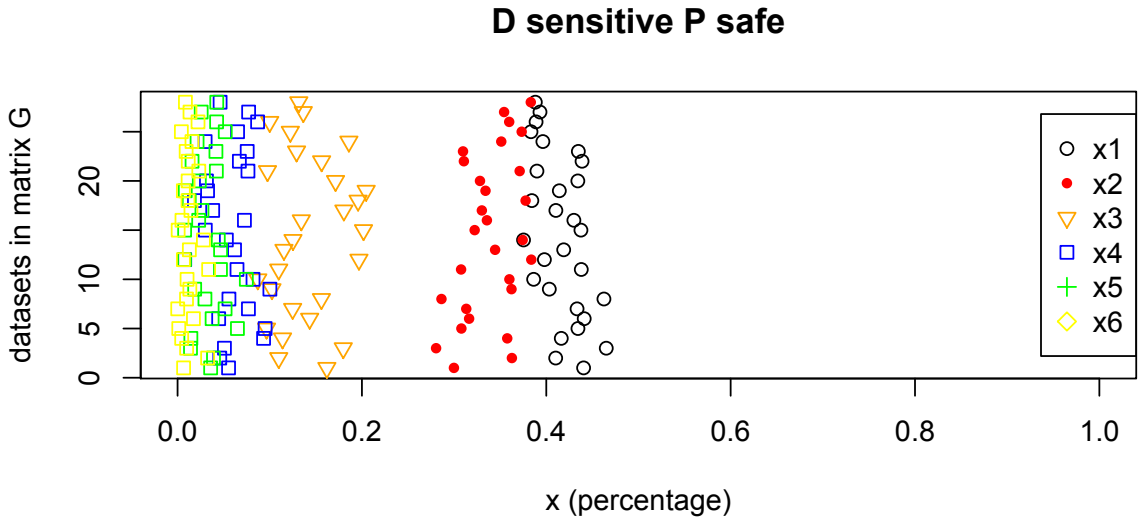
Willenborg, L., and T. De Wall (2001) *Elements of Statistical Disclosure Control*, New York: Springer-Verlag

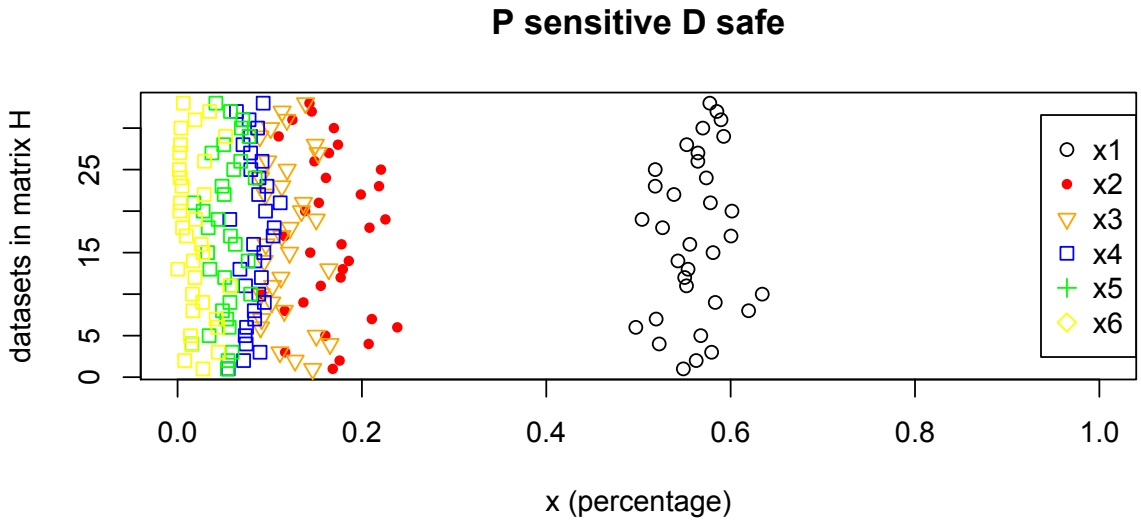
Appendix I Pairwise comparison between rules, when x~exp (λ=1) and n=6

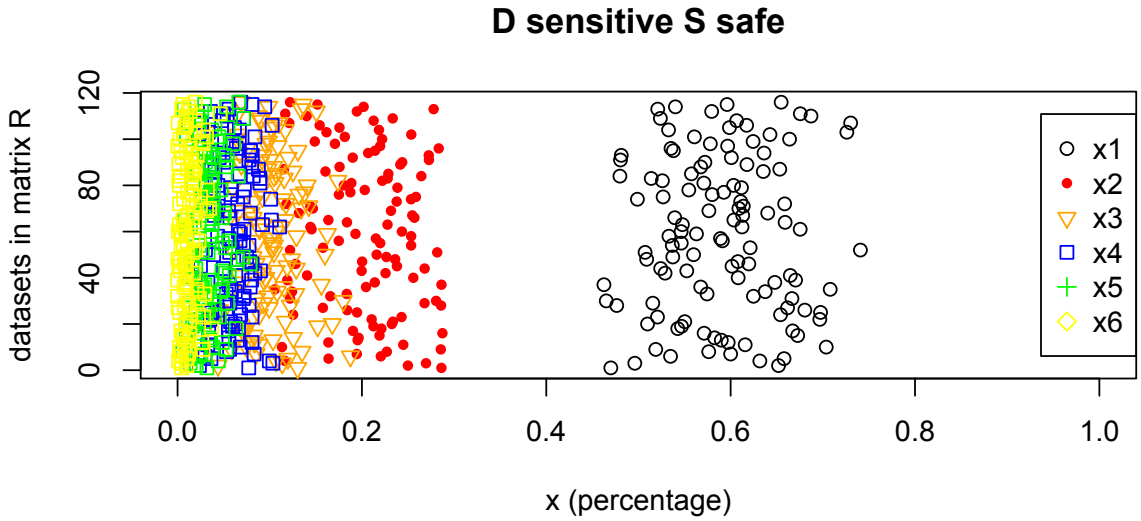


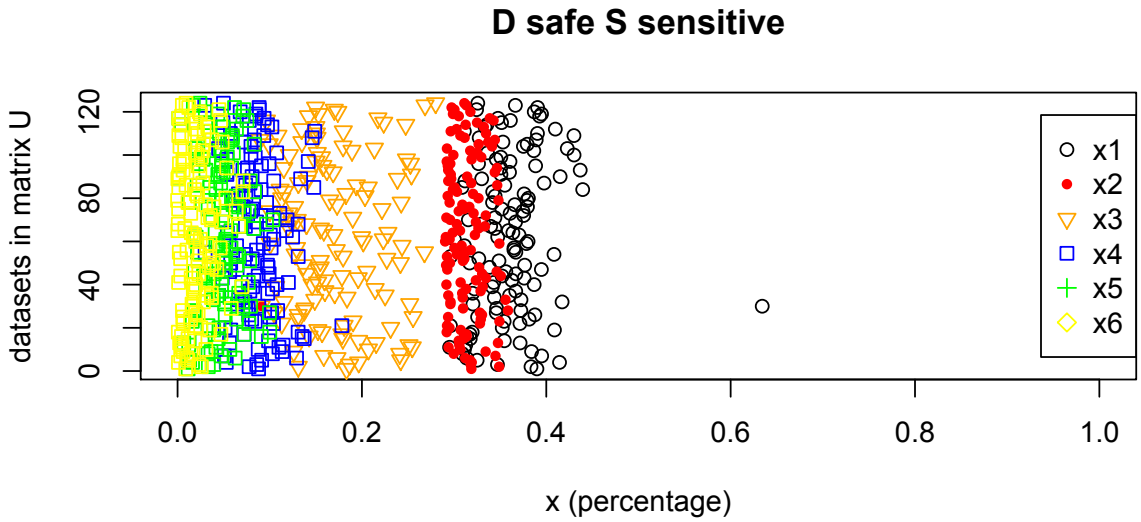

P sensitive S safe
datasets in matrix A
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6


P safe S sensitive
datasets in matrix B
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6

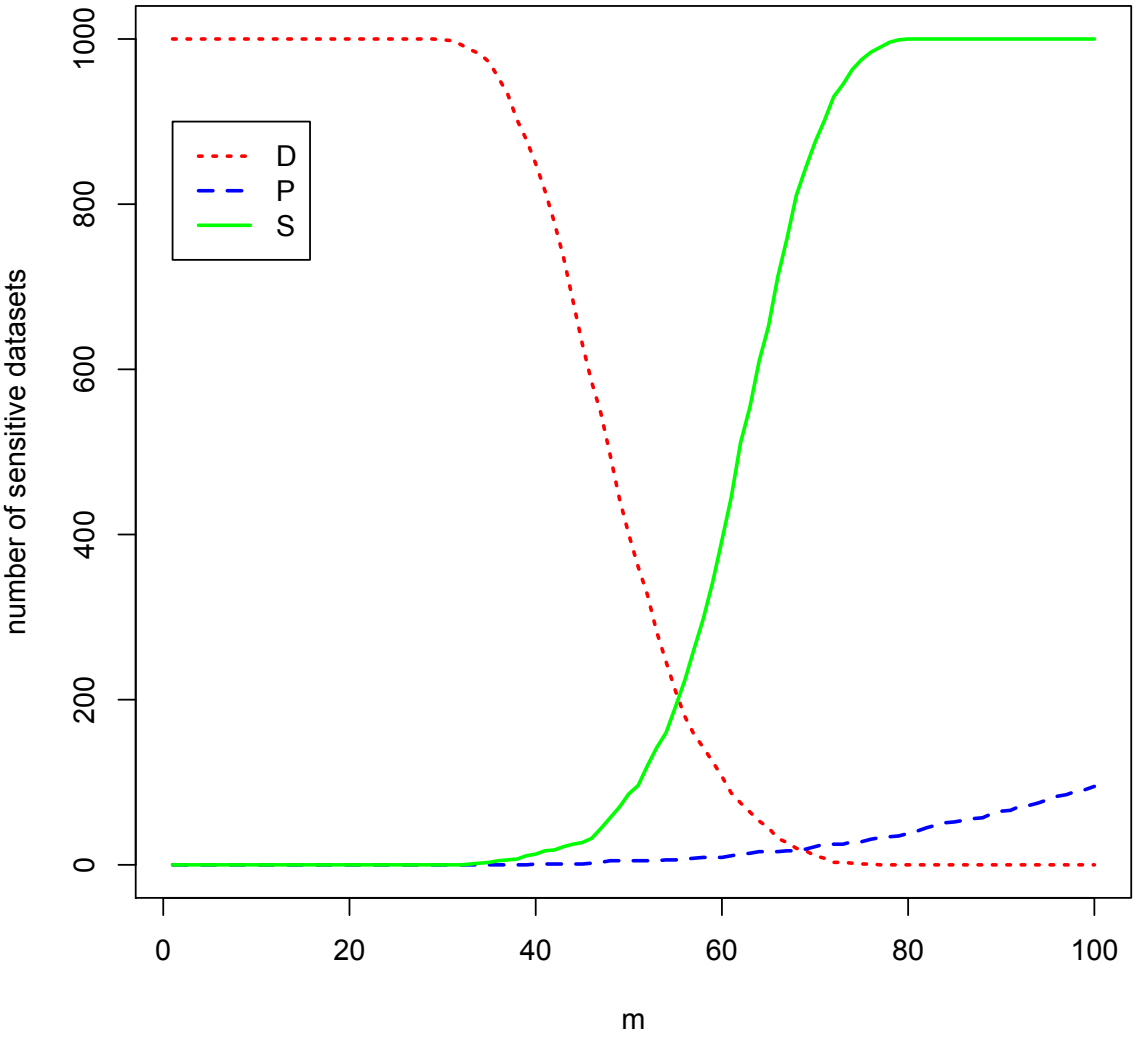

D sensitive P safe
datasets in matrix G
0
5
10
20
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6

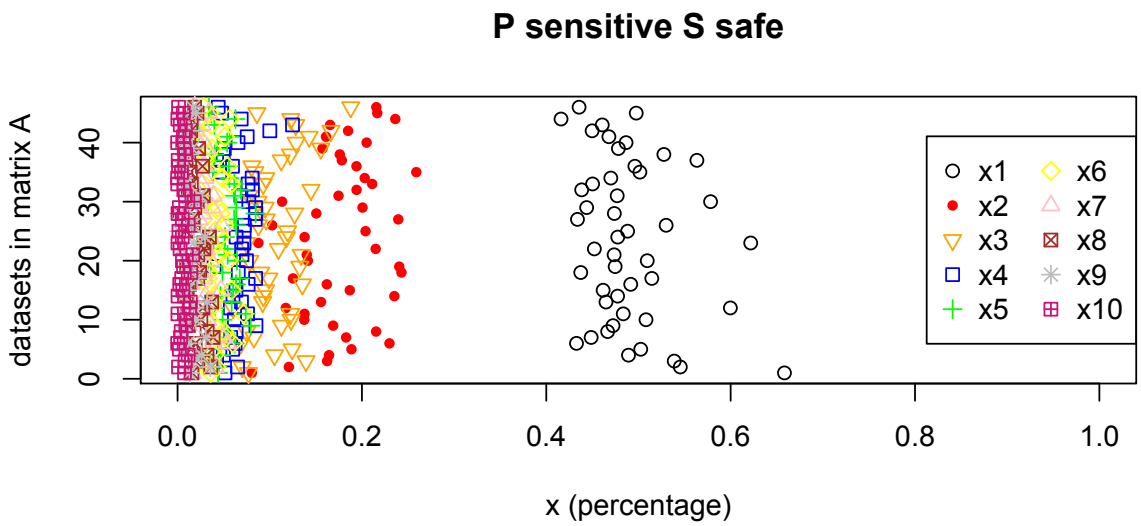

P sensitive D safe
datasets in matrix H
0
5
15
25
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6

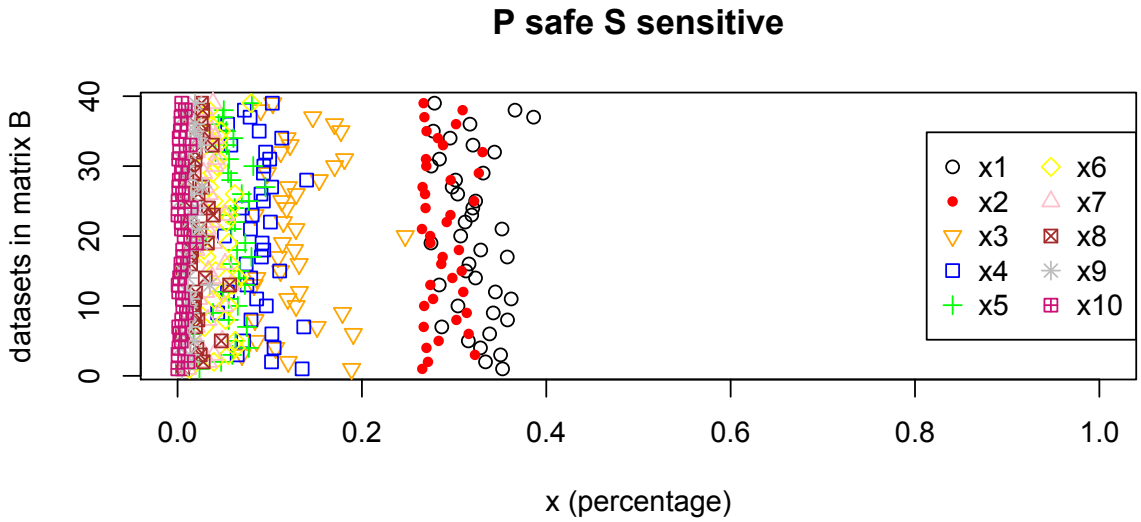

D sensitive S safe
datasets in matrix R
0
40
80
120
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6

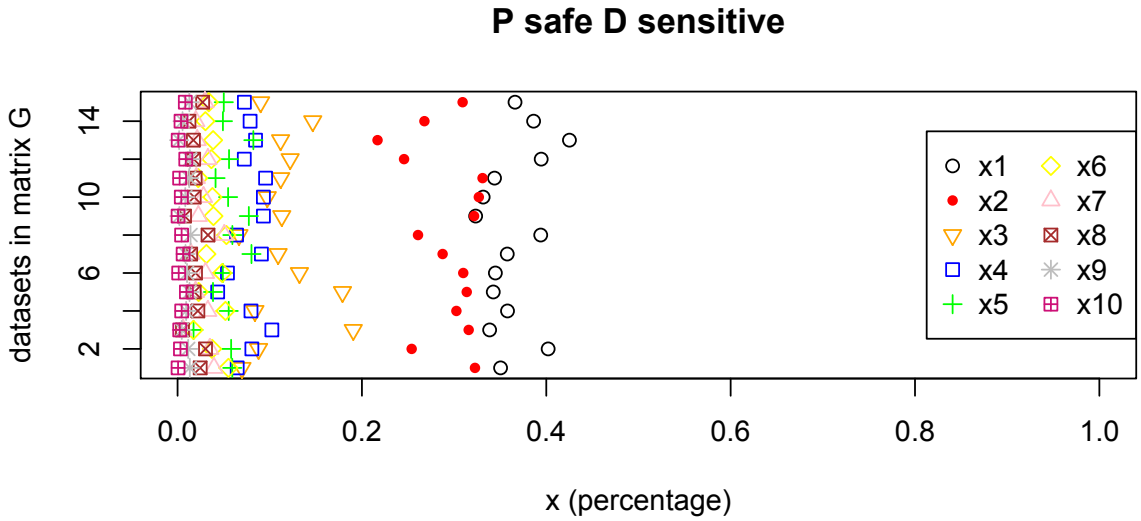

D safe S sensitive
datasets in matrix U
0
40
80
120
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6

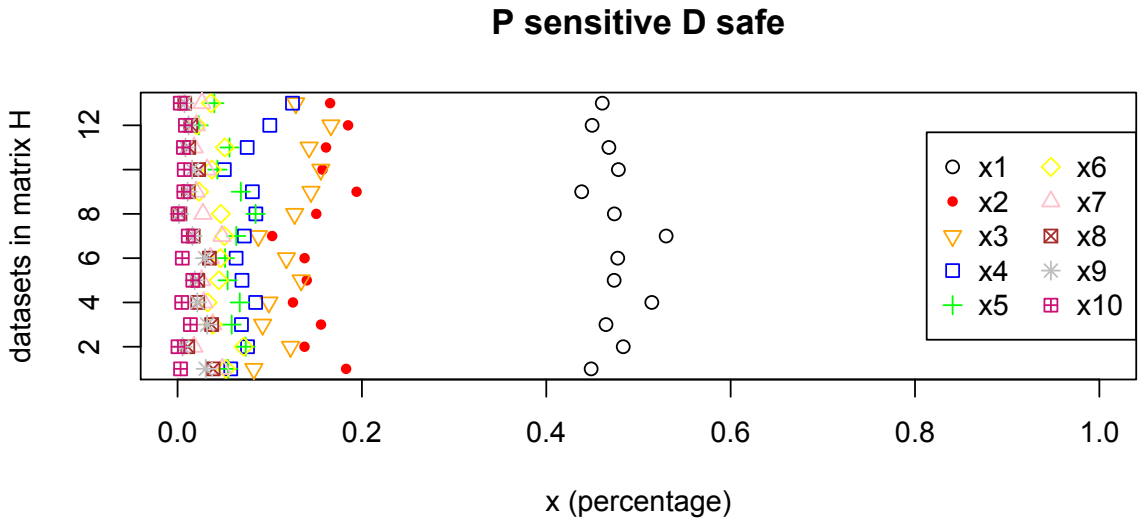
Appendix II Pairwise comparison when x~exp (λ=1) and n=10

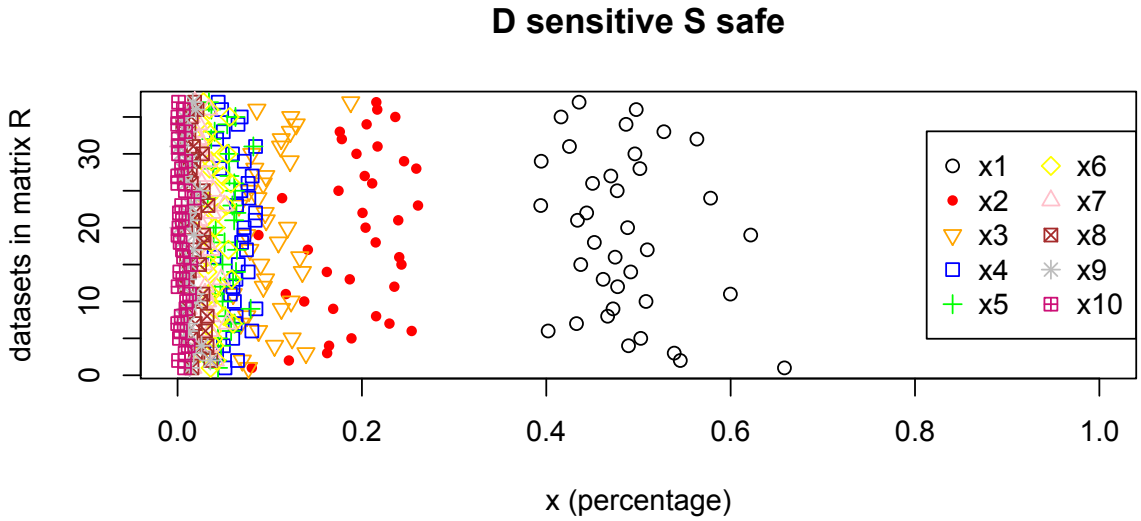


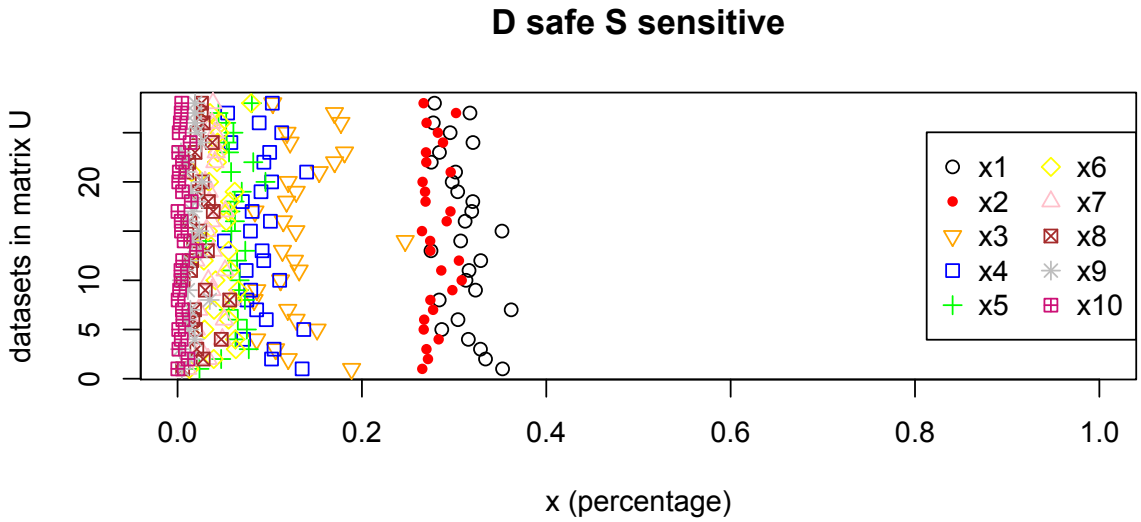

P sensitive S safe
datasets in matrix A
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10


P safe S sensitive
datasets in matrix B
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10

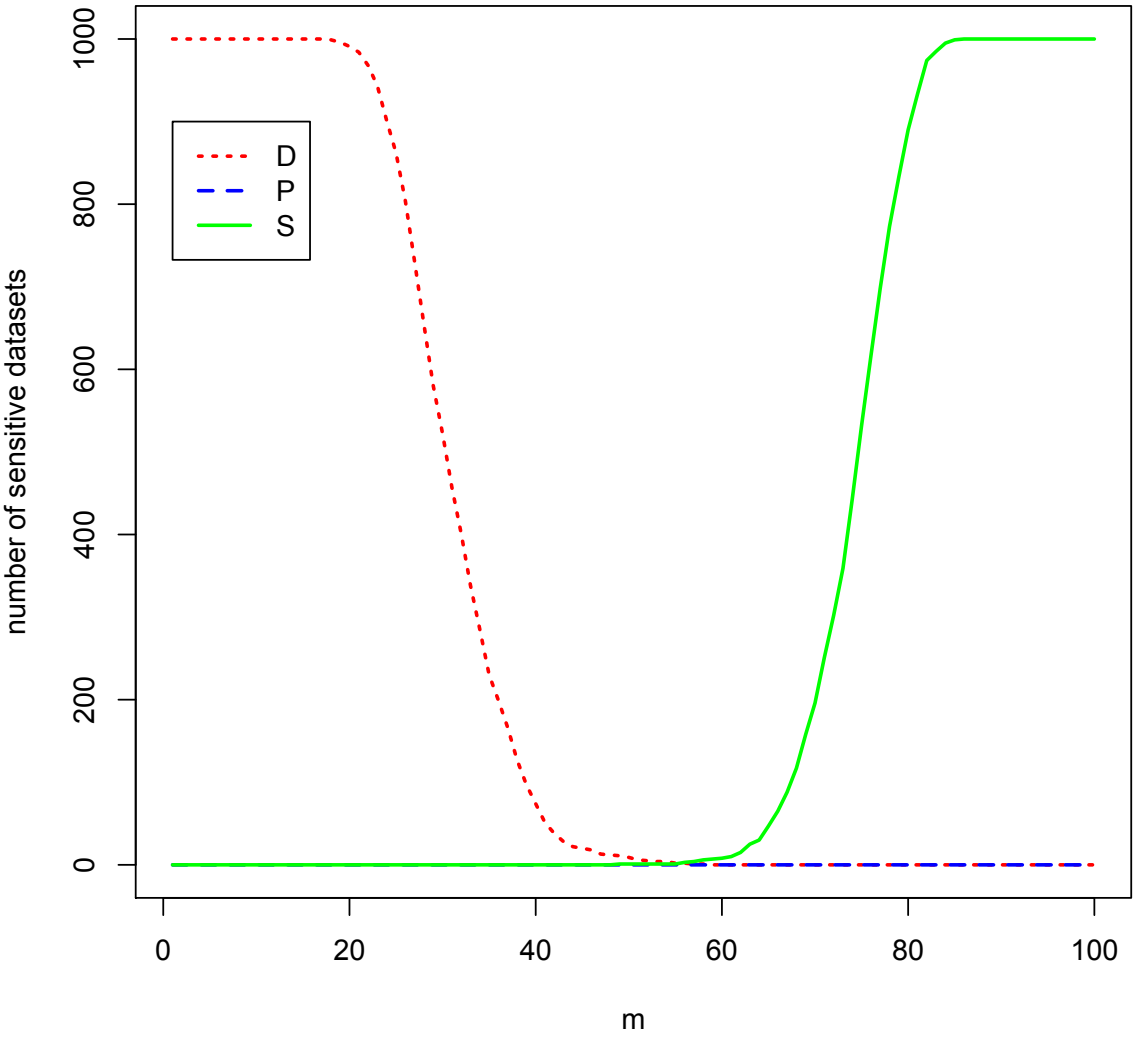

P safe D sensitive
datasets in matrix G
2
6
10
14
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10

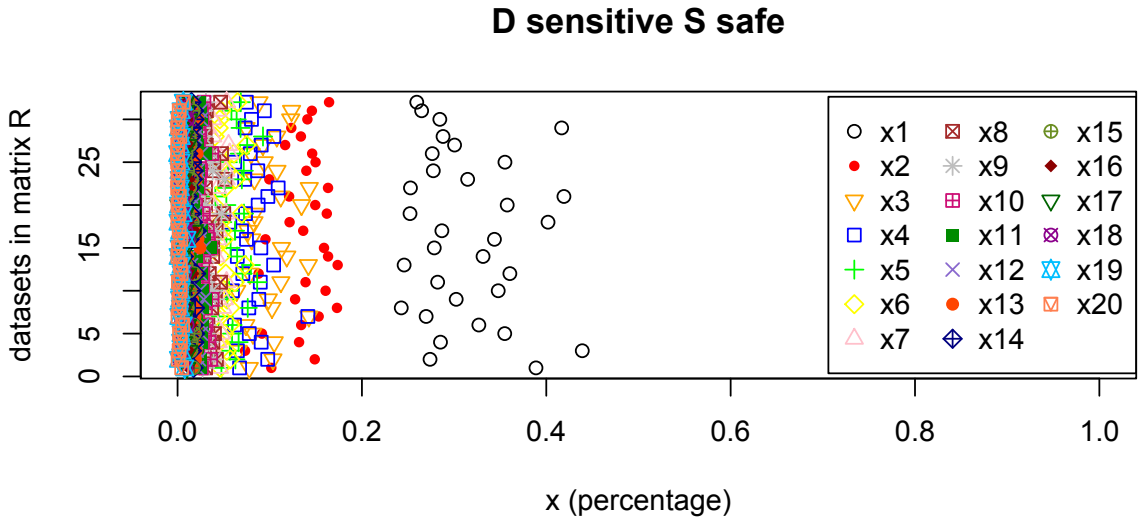

P sensitive D safe
datasets in matrix H
2
4
6
8
12
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10

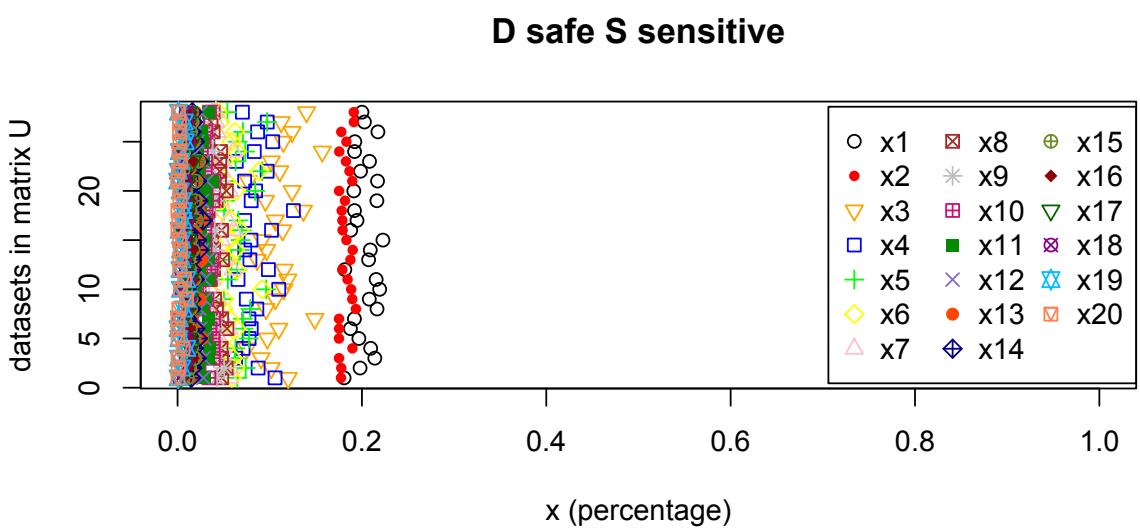

D sensitive S safe
datasets in matrix R
0
10
20
30
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10


D safe S sensitive
datasets in matrix U
0
5
10
20
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10

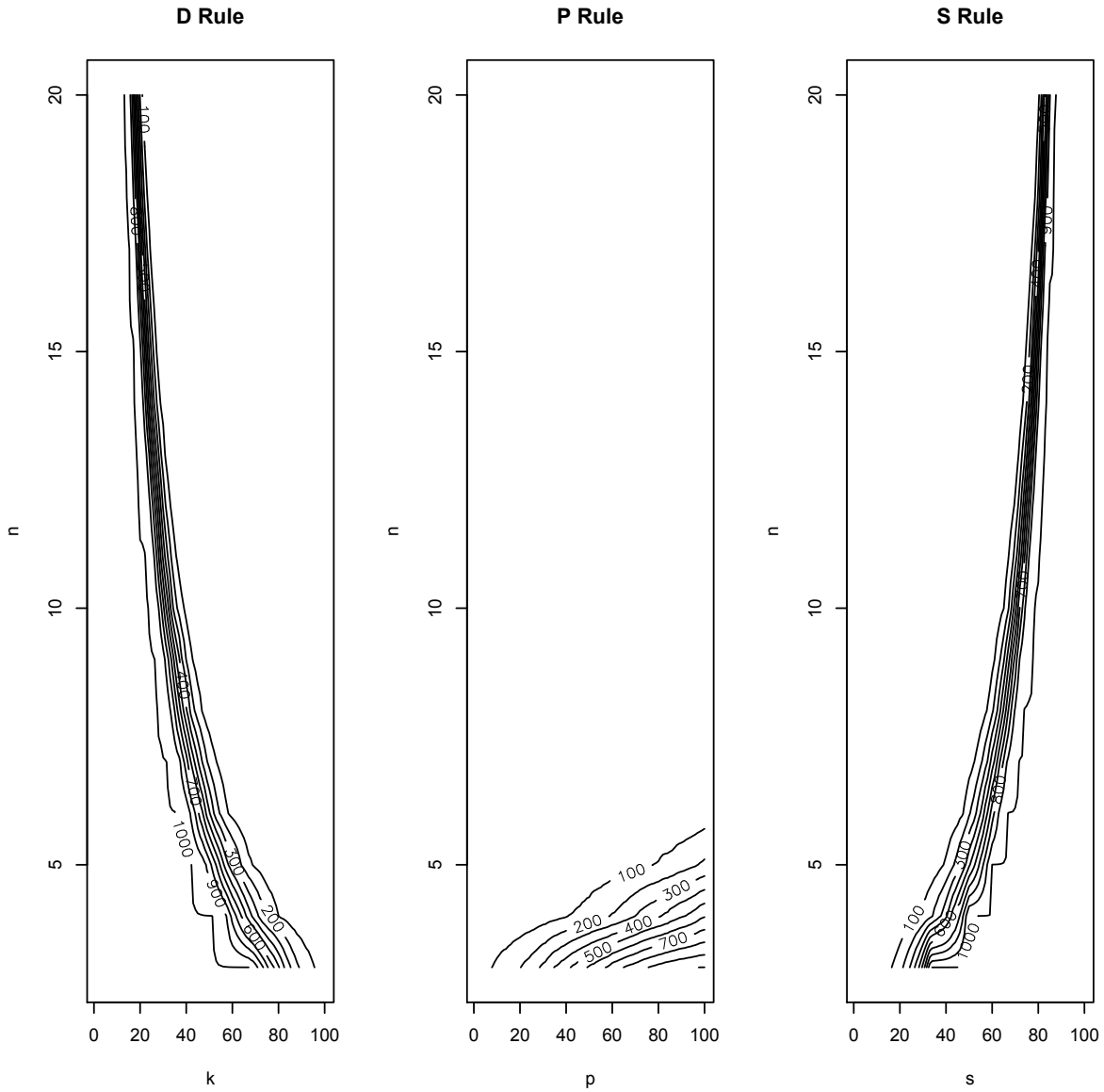
Appendix III Pairwise comparison when x~exp (λ=1) and n=20



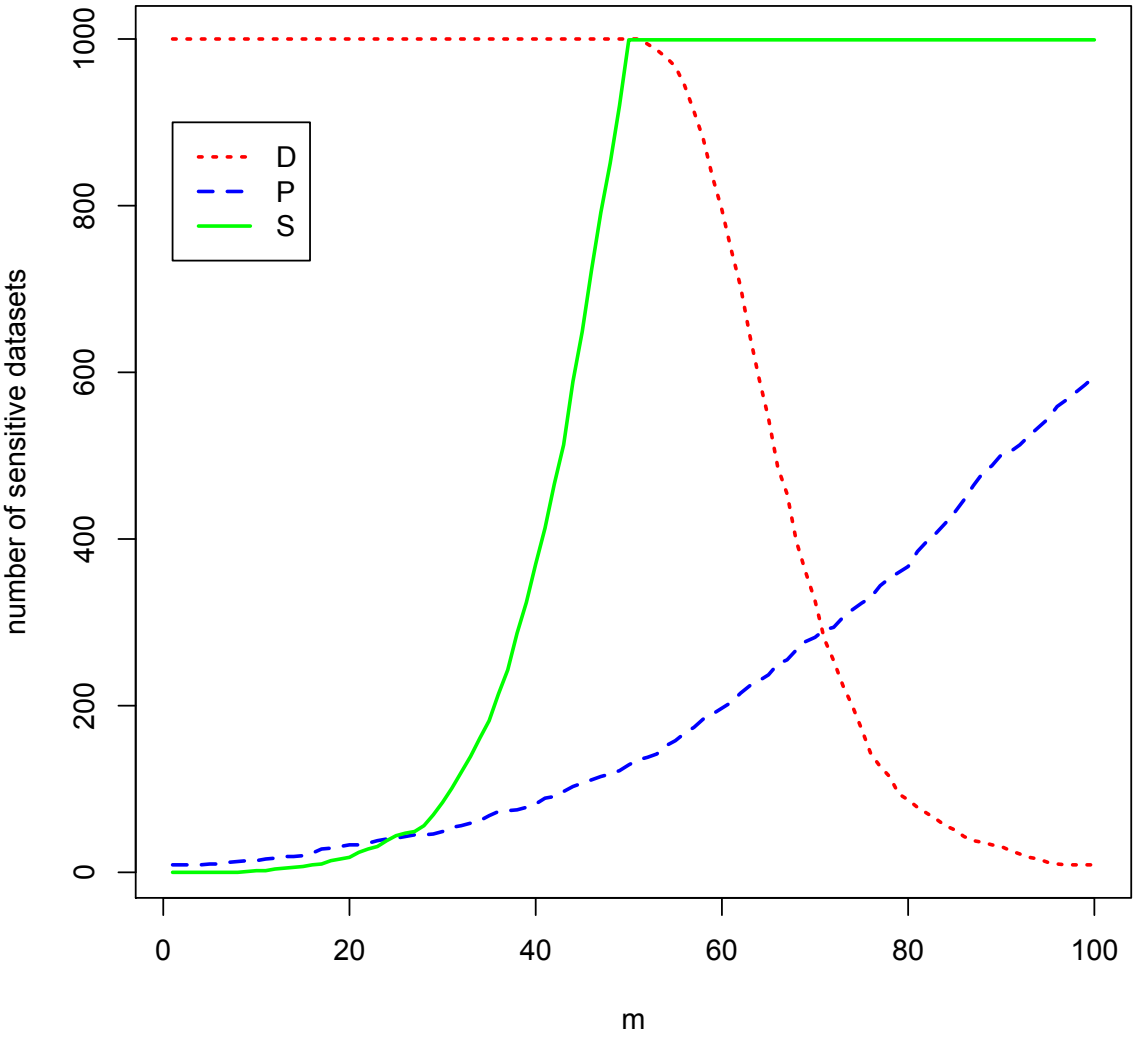

D sensitive S safe
datasets in matrix R
0
5
15
25
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
x11
x12
x13
x14
x15
x16
x17
x18
x19
x20

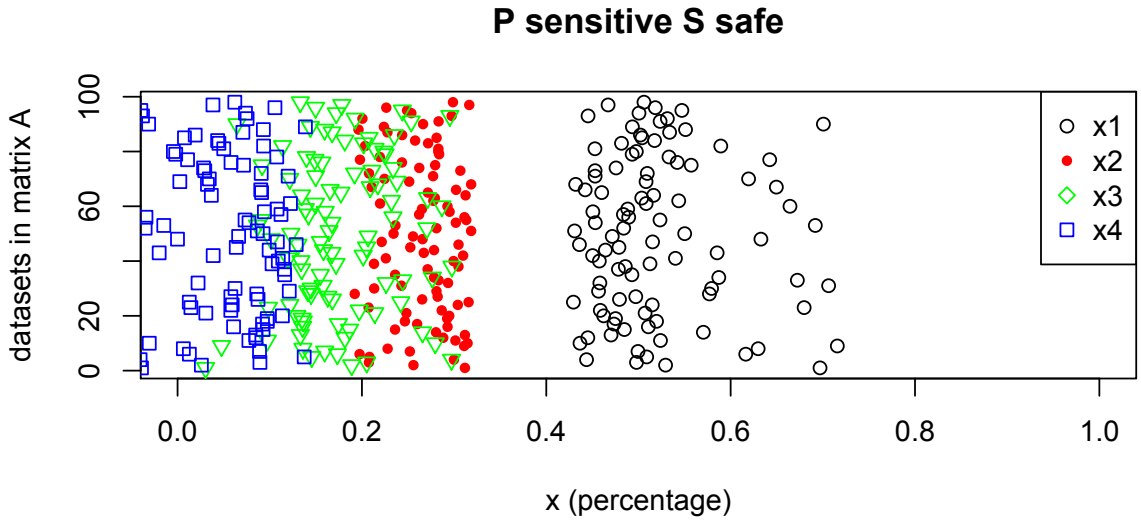

D safe S sensitive
datasets in matrix U
0
5
10
20
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6
x7
x8
x9
x10
x11
x12
x13
x14
x15
x16
x17
x18
x19
x20

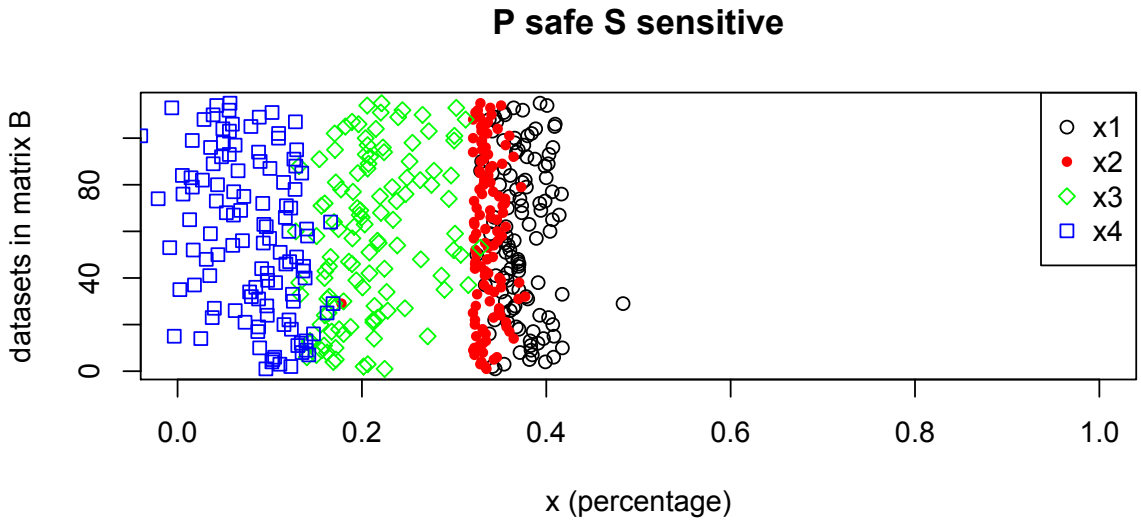
Appendix IV Contour plot when x~N (100,50)

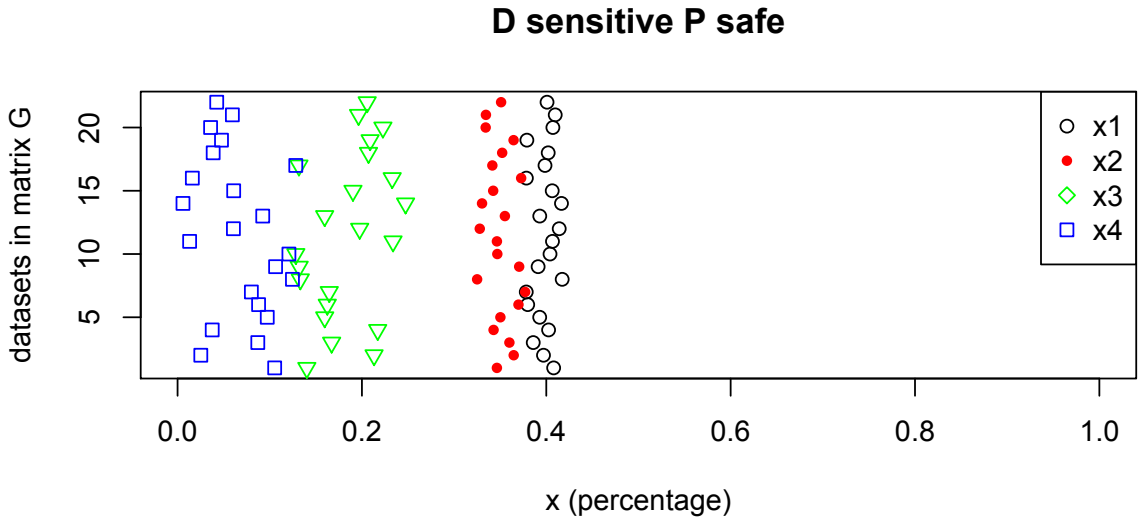


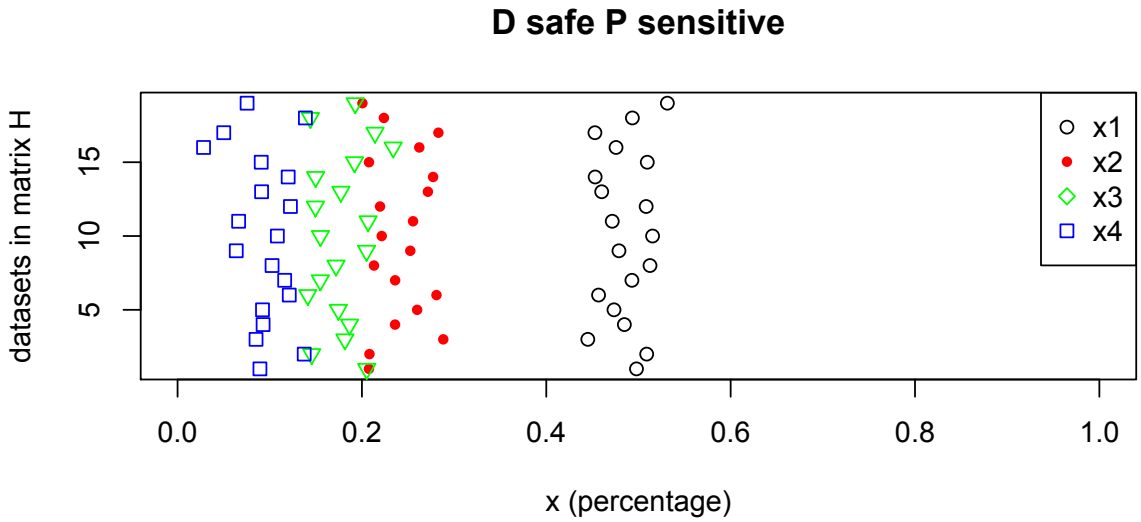
Appendix V Pairwise comparison when x~N (100,50) and n=4

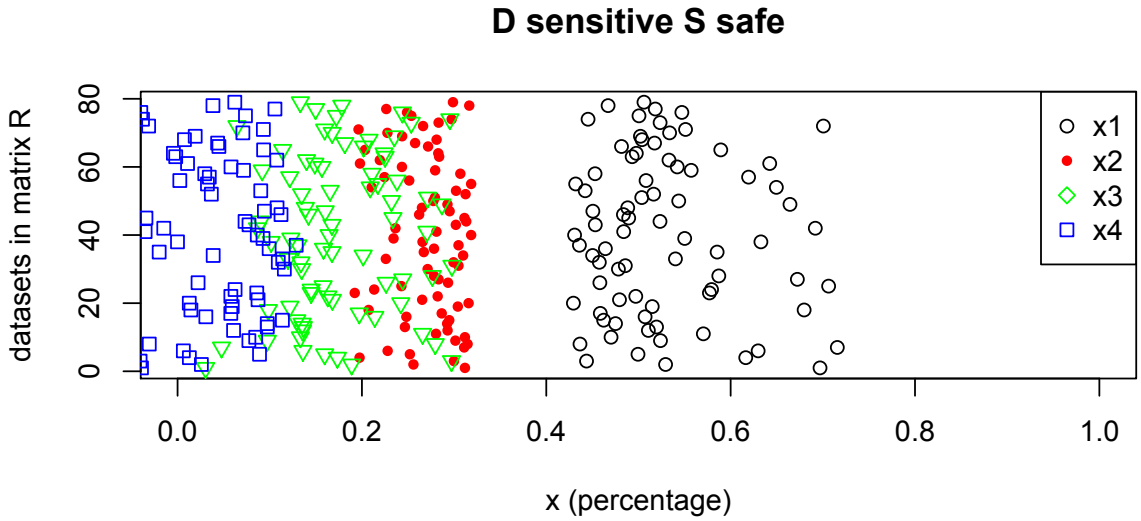


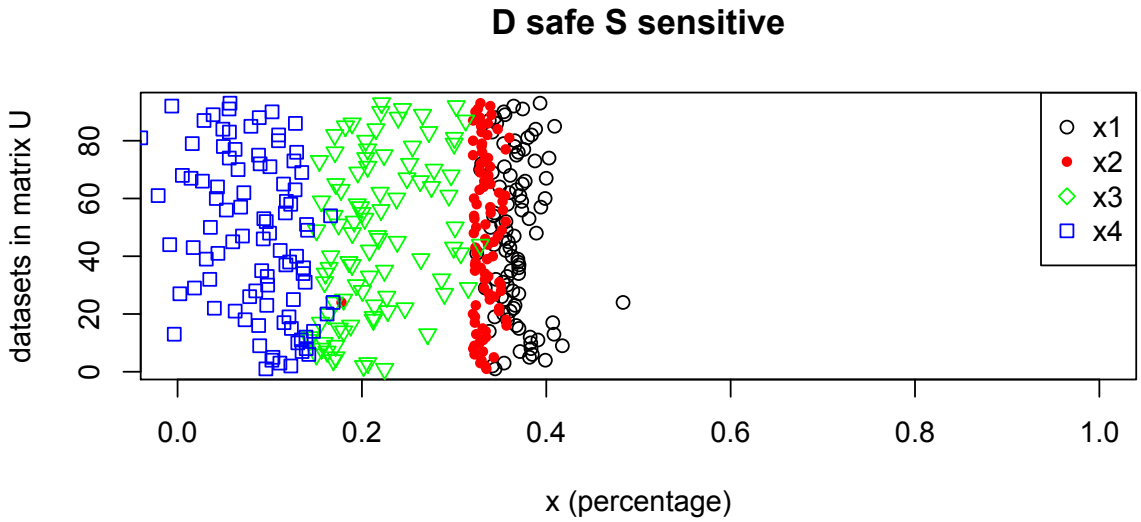

P sensitive S safe
datasets in matrix A
0
20
60
100
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4


P safe S sensitive
datasets in matrix B
0
40
80
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4

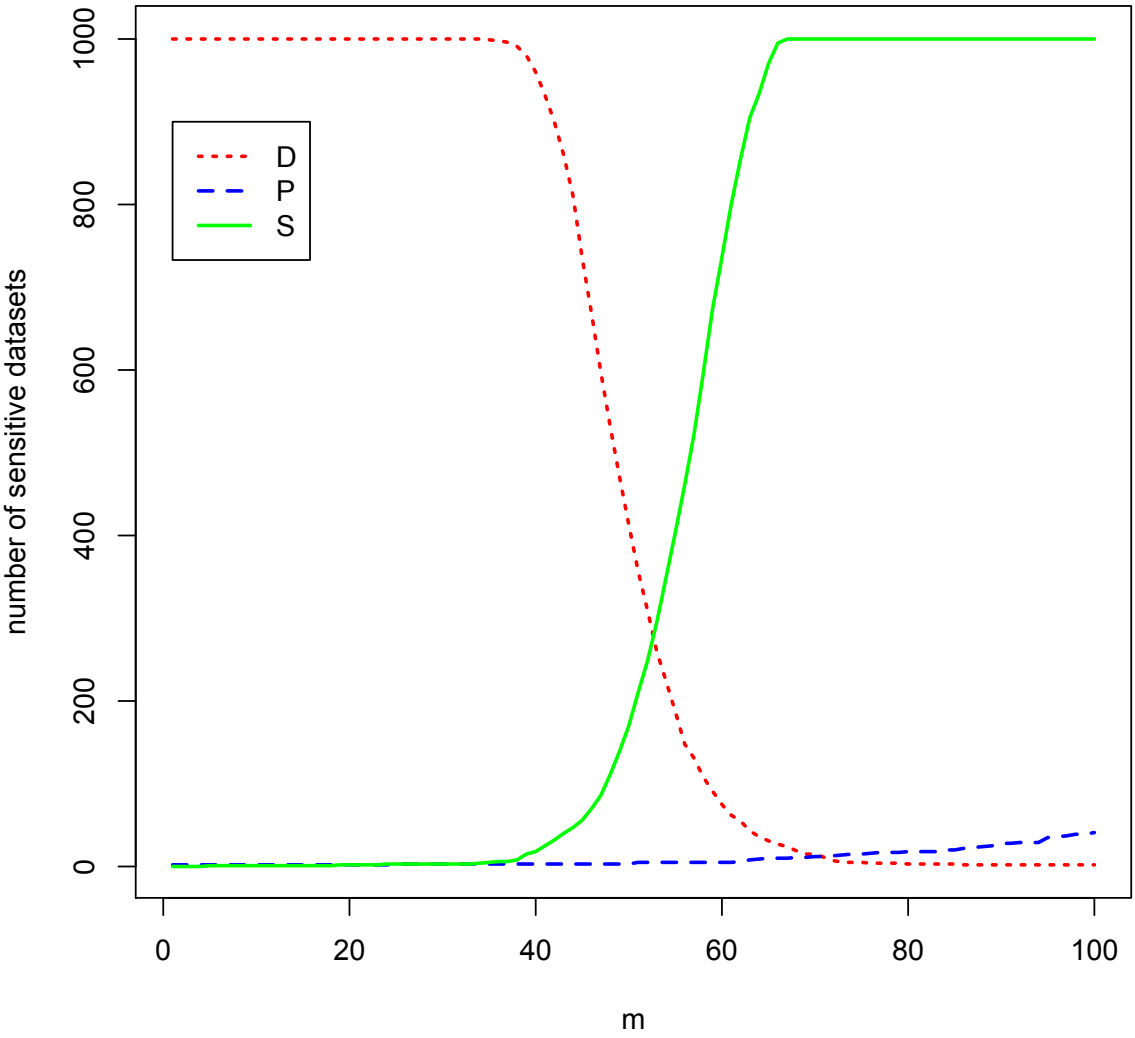

D sensitive P safe
datasets in matrix G
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4

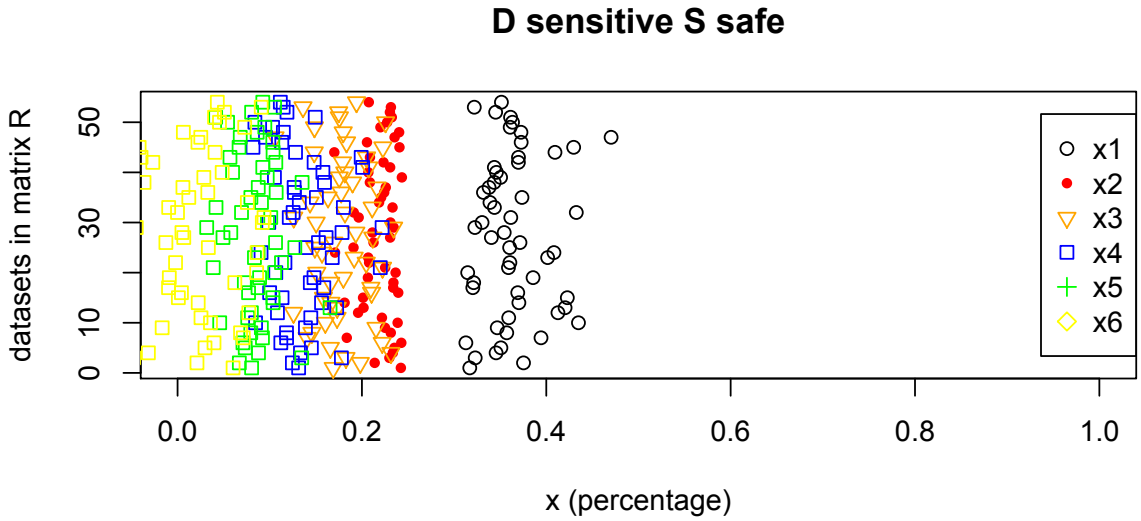

D safe P sensitive
datasets in matrix H
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4

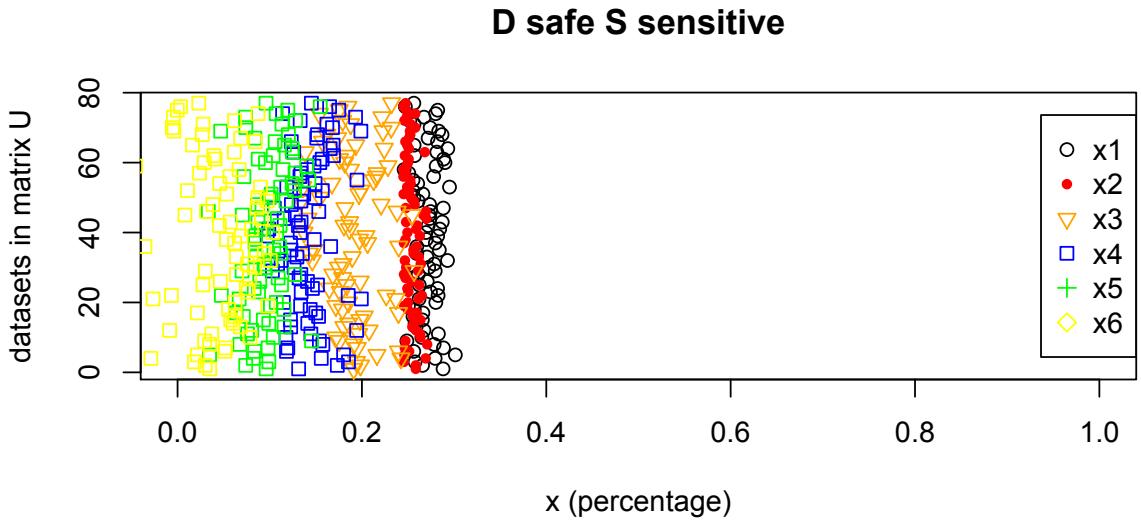

D sensitive S safe
datasets in matrix R
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4


D safe S sensitive
datasets in matrix U
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4

Appendix VI Pairwise comparison when x~N (100,50) and n=6




D sensitive S safe
datasets in matrix R
0
10
30
50
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6


D safe S sensitive
datasets in matrix U
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
x (percentage)
x1
x2
x3
x4
x5
x6