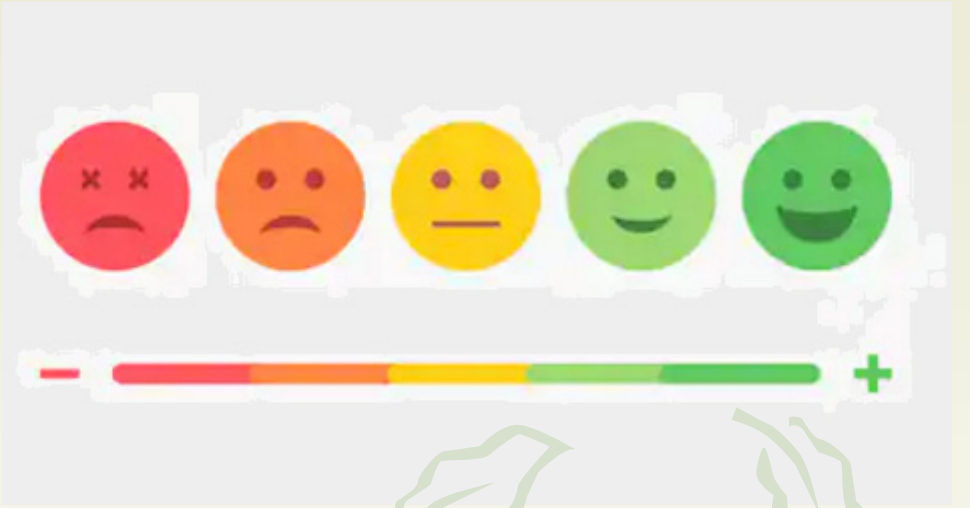


Survey Data Cleaning

SYMPOSIUM I ANVENDT STATISTIK

2006

Informatik og Matematisk Modellering
Danmarks Tekniske Universitet

Danmarks Statistik

Survey Data Cleaning *

Joseph E. Mullat, Independent researcher
Copenhagen, Denmark, mailto: mjoosep@gmail.com;
Former docent at the Faculty of Economics, Tallinn Technical University, Estonia

Abstract.

The note addresses a data cleaning principle. The principle implementation procedure presented here includes a recommendation that might be well suited for explicating and illustrating the results yielded by survey data analysis.
Keywords: Clustering, Data cleaning, Dirty data, Customer Satisfaction

1. INTRODUCTION

Every day, in an endless stream, we are presented with various polls, studies, statistics, opinions, measurements, research results, etc. Enterprises, media experts, universities and other interested organizations try to present reality in a certain way or explain how it all works using information in the form of data collected during the interview. While we take this influx of data for granted, very few of us question whether this way of having reality served on a platter is actually helpful. Most people merely accept what the various analysts have presented and treat it as factual information. Thus, if more people in a survey have answered that they prefer rye bread to the white variety, does the same assertion apply to the world population? Should we infer from this finding that people in general eat more rye bread instead of white? Certainly not, reality is complex and consists of numerous choices, possibilities, behavioral patterns, preferences, etc. As a result, a typical survey based on which such 'facts' are reported can never cover all relevant data pertaining to any given subject and would without doubt lead to completely nonsensical conclusions. More accurate approximations of reality require a comprehensive statistical investigation. Therefore, as a rule, when aiming to interpret data gathered based on a sample drawn from a population of interest, one should seek input from a researcher or some other qualified person, so that the

* Presented at the 19th Nordic Conference on Mathematical Statistics, June 9-13, **2002**, Stockholm, Sweden and at the "Symposium i Advent Statistik," January 23-26, **2006**, København.

results can be interpreted and analyzed. Additionally, it is essential to take into consideration the researcher's knowledge and expertise on the subject, as well as carefully assess whether the questions discussed pertain to the aim of the survey. It is equally important to evaluate the respondents' credibility and ability to answer the questions posed, as this is one of the means to ensure the instrument reliability.

2. Reliability

Reliability, as a generic concept, is difficult to define. In most cases, it is interpreted in a specific context. Nevertheless, it can be shows that adopting the "maximum principle" will not only help the researcher in his/her analytical endeavors, but will also "clean up" the investigation, filtering out the more "unreliable" answers and thus remove some "interference" or "outliers"—i.e. answers that are overly dissimilar from the rest or are incongruent with the most conceivable result. However, it must be emphasized that the method of analysis is still central to the success of the outcome. In other words, in spite of the aforementioned argument, the final estimation should still be based on the subjective perception of reality. After all, the primary difference between this method and the conventional statistical analysis employed to interpret survey results is that the former identifies both *unreliable* respondents and their *unreliable* answers. Consequently, we hereby obtain a much more comprehensive picture of reality simply by examining patterns that conform to the answers provided by the remaining group members. In order to describe the method, an example of a survey in progress, not having a serious purpose or value, will be used. It should be noted that what follows is significantly simplified, as the main objective is to outline the foundations of the method.

Food is a subject of public interest and related data is thus frequently under the analyst's scrutiny. Hence, in our hypothetical or frivolous example, the objective is to map people's taste preferences. To do so, the survey respondents are presented with five menus listed below and are asked to state their daily consumption of each of the given food groups.

The options they are given are as follows:

1. *Dairy produce: cheese and milk*
2. *Cereals: bread, potatoes, rise and pasta*
3. *Vegetables: vegetables, fruit, etc.*
4. *Fish: shrimp, frozen/fresh fish*
5. *Meat products: various meats, sandwich spreads and sausages*

The results pertaining to seven study participants are presented in Table 1, which will suffice for the upcoming food preferences investigation.

Table 1.

	Dairy	Cereal	Vegetables	Fish	Meat	Total
Respond. no. 1		X	X			2
Respond. no. 2	X	X		X	X	4
Respond. no. 3			X	X		2
Respond. no. 4	X	X		X	X	4
Respond. no. 5			X	X		2
Respond no. 6	X	X	X	X	X	5
Respond. no. 7		X	X			2
Total	3	5	5	5	3	21

Considering the total score given at the bottom of the table, people's food choices seem healthy and nutritional. Moreover, it can be discerned that "cereals," "vegetables" and "fish" are most frequently consumed food groups, as five of seven respondents stated that they consume these foodstuffs daily. Can we conclude that, in general, people's lifestyle is healthy? Moreover, does this mean that 71% of population eats cereals, fish and vegetables every day? This conclusion could be clearly misleading. In addition, even conclusions pertaining to this small group require close examination of the individual respondents' answers, because some of them differ from those of the other respondents in certain ways. For example, respondents 1, 3, 5 and 7 have chosen only two food groups from the given list. Respondents no. 1 and 7 stated that they consume only "cereals" and "vegetable" products on a daily basis, while no. 3 and 5 eat only

"vegetables" and "fish" every day. Assuming that this is an exhaustive list (again, note the simplifications in this example), it seems highly unlikely that someone would not eat any products from other food groups. This is a crucial point to consider, as we must believe that the answers respondents provide and factual in order to include them in the analysis. Thus, responses like those noted above are clearly unreliable reflections of reality. Let us therefore experimentally discard the unreliable respondents together with their answers to see whether we obtain a more credible result, which is a more accurate representation of reality.

3. Agreement Level – Tuning Parameter

Just as it is unusual to rely on only two food groups for sustenance, it is unlikely that an individual would eat, for example, only bread from the cereal menu, or solely shrimp from the fish menu. Thus, in "fine-tuning" the experiment, the aim is to identify all the respondents that have chosen only these two menus. The objective is, as was already emphasized above, to obtain a clearer picture of reality. Table 2 below represents the results of this data "cleaning," based on the chosen "agreement level" or "tuning parameter". In this case, the agreement level is set to 4, i.e. none of the totals in the last column is less than 4.

Table 2.

	Dairy	Cereal	Vegetables	Fish	Meat	Total
Respond. no. 2	X	X		X	X	4
Respond. no. 4	X	X		X	X	4
Respond. no. 6	X	X	X	X	X	5
Total	3	3	1	3	3	13

This seems to be a very useful instrument for the experiment. However, the tuning parameter will only be relevant when its value exceeds one. If, for example, we try to set the agreement level (tuning) to 1 in Table 1, this would render ALL respondents reliable, even though menus "Dairy" and "Meat" are associated with the lowest frequency number, namely three.

What can we conclude from the outcome of adopting tuning parameter = 1? The conclusion is exactly the same as that yielded by the original analysis—"people's lifestyle is healthy." In contrast, setting the tuning parameter to 2, 3 or a higher value allows us to explore patterns in answers that would not be otherwise apparent. Table 2 shows the distribution of respondents based on the tuning parameter = 4.

Why should we use this particular value as a tuning parameter? Yes, indeed, in the following analysis we intend to adopt the maximum principle as a method for selecting reliable respondents. This will be done through "agreement level", see "*totals*" of columns, pertaining to a single respondent. The value of the tuning parameter is not fixed, and can be changed depending on the purpose of analysis, and is typically set at the level that reveals the most adequate picture of reality. Roughly speaking, we can compare the situation to rotating a tuner on TV or Radio, when we attempt to receive a clear picture/sound by trying to select the right frequency. The tuner value here is 4, and we assume that the selected respondents are now reliable.

4. MAXIMUM PRINCIPLE

Finding the correct tuner position is not sufficient, as will be shown in the discussion that follows. For example, only one of the remaining, supposedly reliable, respondents chose the "vegetable" menu. This would imply that only 33% of the sample is consuming vegetables daily. While this is likely for such a small group of respondents, it is important to reiterate that this example is a simplification of an actual, much larger survey, where such results would indeed be odd. Thus, the fine-tuning must proceed further, this time addressing the menu content. Fist, we can remove "vegetables" from the available options and see what effect this would have on the analysis.

The next step in our analysis is called "maximum principle" (Mullat, 1971a) and will be illustrated using an old merchant marketing example. If a merchant wants to make a compromise between the highest possible demand on some assortment of his/her commodities and to shorten the

list of assortments as well, he would intuitively do so by removing from the assortment the commodity for which the demand is the lowest, assuming that it is identified from the purchasing patterns of reliable customers only. In the example considered in this study, the "vegetables" menu has the lowest demand. Moreover, its removal from the available options results in equal frequencies associated with the remaining menus. In general, removal of available options must be done with care, as it should not result in a simultaneous removal of reliable respondents. In some cases, however, it might be necessary to add further reliable respondents to the sample, complying with our tuning parameter once again, etc.

In general, the maximum principle can be formulated as follows: among all the reliable respondents, first remove options with the lowest agreement level, those with the lowest frequency (in our example, the menu "vegetables" in Table 2). As a result, the number of choices is reduced, but the remaining answers with the lowest frequency have a higher contingency compared to those that have been removed. In short, the aim is to remove available options in such a manner that ensures that those remaining have high representation and there are more matches in their answers. In other words, in the menu, where the matching is low, the low match becomes relatively high due to the removal, which would not be the case if the removed menus will still occupy a place in the table. In other words, the goal is not only to separate a group of menus from those that have higher matching responses, but also to find a group of respondents for whom the menu with the lowest level of matching is on a relative high level. This is the key for understanding the maximum principle. The respondents included in the analysis must be reliable, but the answers producing such reliability must also be more or less identical.

In accordance with this argument, the menu "vegetables" is removed, since the responses associated with it are not aligned with the general answer pattern based on the maximum principle. Note that here, the removal is not based on any qualitative tests, but is rather guided purely by a pattern disclosed by matching the answers!

Table 3.

	Dairy	Grain	Fish	Meat	Total
Respond. no. 2	X	X	X	X	4
Respond. no. 4	X	X	X	X	4
Respond. no. 6	X	X	X	X	4
Total	3	3	3	3	12

5. CONCLUSION

What can be concluded from the simplified survey scenario discussed above? Put it simply: it is evident that the final outcome is completely different from the results yielded by the initial analysis. According to Table 1, in general, people's food preferences are healthy and in accordance with current recommendations. On the other hand, Table 3 indicates that food habits are, in fact, less healthy. Implementing our analysis principle has reduced the panel of reliable respondents, and this has changed the outcome of our analysis.

Of course, it is natural to ask whether the proposed principle is more credible than other methods of analysis. It is true that a subjective consideration and personal choice have played in instrumental role in the analytical framework adopted to produce the final results. Some may argue that this approach is flawed, as analyst/researcher intuition was the only basis for tuning the parameters, i.e. adjusting the "agreement level." This personal consideration cannot be excluded because the method described here will sometimes coincide with what we might otherwise call common sense, where the most frequent answers reflect the actual reality. This should be the case when dealing with simple surveys in which the respondents are asked questions such as "Will you vote for so and so the coming election?" The value of this approach is really evident when surveys including hundreds or thousands of respondents and many hundreds of questions are conducted. They will inevitably generate diverse responses forming patterns that "common sense" will be impossible to wield, since unaided human intellect is incapable of grasping such com-

plicated patterns. This is where our method can make a substantial difference, because it is a way of locating erroneous or misleading patterns, based on a comprehensive comparison within the full data set. This, however, does not undermine the analysts' role, as these experts will be responsible for making the relevant judgments/decisions as to why certain data is removed from the set. The goal is to identify and remove all "unreliable" respondents with the help of the "tuning parameter." The aim of this "cleansing procedure" is to retain only the most usable answers, in accordance with our maximum principle. Thus, the method presented here should be treated as an instrument, which has to be used correctly by the analyst to tune into the clearest picture of reality. The aim is to reduce the interference effect produced by unreliable respondents.

APPENDIX

A.1 Practical recommendations

The preliminary explanation above is a general introduction to our maximum principle, the background of which is found in a much more complex methodology and theory.[1] First, it is beneficial to demonstrate how the results can be used and presented for the analyst, making the use of the notion of positive/negative profile.

When designing a questionnaire, it is widely accepted that the available responses associated with the individual questions should be presented in the "same direction," i.e. from positive to negative values/opinions or vice versa. Using a more rigorous terminology, such ordering would be denoted numerically and represented on an nominal/ordinal scale. This nomenclature is used primarily because, when implementing our method in the form of computer software, the analyst must separate the answers by grouping them together into positive/negative scale ends—the $(+/-)$ pools. The next step will be to create profile groups within each $(+)$ or $(-)$

[1] Some theoretical aspects may be found in Appendix A.2

pool range. A profile group of answers is created following their subject-oriented field of interest. For example, one might be interested in participants' lifestyle, nutritional practices, exercising, etc. Thus, these profiles, distinguished by their placement in $(+/-)$ pools, are also either positive or negative.

Once the analyst has created the $(+/-)$profiles, an automated process utilizing our maximum principle, which further organizes the data into what we call a series of profile components, conducts the subsequent analysis. Each profile component is a table, as above, located within particular profile limits. Clearly, a component is differentiated from the profile by the fact that, while a profile is a list of subject-specific questions and the corresponding options/answers composed by the analyst, the component is a table formed using the maximum principle. Therefore, the list of answers constituting a component (and the resulting set of table columns) is smaller, as only specific answers/columns from the full profile are included. Thus, once again the components will be separated into $(+/-)$ components $K_1^{\pm}, K_2^{\pm},...$, just as the profiles were separated into $(+/-)$ profiles. The $K_1^{\pm}, K_2^{\pm},...$ separation provides not only conceptual advantages, but also allows for more transparent illustration of the survey findings.

Analysis findings increase in value if they are presented in the format that can be easily comprehended. The simplest tool available for graphical presentation is a pie chart. Here, the pie can be divided into positive $K_1^{+}, K_2^{+},...$, and negative $K_1^{-}, K_2^{-},...$ components, represented in green and red color, respectively. However, to depict these components accurately, it is necessary to calculate some statistical parameters beforehand. For example, one can merge the $(+/-)$ components into single $(+/-)$ table and calculate the $(+/-)$ probabilities.[2] Hereby, statistical parameters based on the $(+/-)$ probabilities may be evaluated and illustrated by a pie chart

[2] Certainly, some estimates only.

divided into green and red area, effectively representing the $(+/-)$ elements.[3] There are many techniques and graphical tools at the analyst's disposal, and a creative analyst may proceed in this direction indefinitely. Still, it is plausible to wonder if the creation of the $(+/-)$ components is worthwhile. In other words, what is the advantage of using the "maximum principle" when interpreting the survey findings? The answer, see above, is that the blurred nature of the data may hinder clear interpretation of the reality underlying the data.

A.2 Some theoretical aspects

Suppose that respondents $N = \{1,...,i,...n\}$ participate in the survey. Let x, $x \in 2^N$, denote those who expressed their preferences towards certain questions $M = \{1,..., j,..., m\}$. We lose no generality in treating the list M as at a profile, whether negative or positive. Let a Boolean table $W = \left\| a_{i,j} \right\|_n^m$ reflect the survey results related to respondents' preferences, whereby $a_{i,j} = 1$ if respondent i prefers the answer j, $a_{i,j} = 0$ otherwise. In addition, all lists 2^M of answers $y \in 2^M$ within the profile M have been examined. Let an index $\delta_{i,j}^k = 0,\ i \in x, j \in y$ if $\sum_{j \in y} a_{i,j} < k$, otherwise $\delta_{i,j}^k = 1$, e.g. $\sum_{j \in y} a_{i,j} \geq k$, where k is our tuning parameter. We can calculate an indicator $F_k(H)$, using sub-table H formed by crossing entries of the rows x and columns y in the original table W. The number of 1-entries $\delta_{i,j}^k \cdot a_{i,j} = 1$ in each column within the range y determines the indicator $F_k(H)$ by further selection of a column with the minimum number $F_k(H)$ from the list y.

---

[3] Please, find below a typical pie chart pertinent to what we just discussed. The positive and negative profiles relate to 21 questions highlighting people's behaviour, responses, opinions, etc., regarding their daily work and habits. Answers to these questions can be presented using an ordinal scale 1, 2, ..., 5, where 1, 2, 3 are at the negative, and 3, 4, 5 at the positive end of the scale.

Identification of the component K seems to be a tautological issue, in the sense that following our maximum principle we have to solve the indicator maximization problem $K = \arg\max_{(x,y)} F_k(H)$. The task thus becomes an NP-hard problem, the solution of which includes operations that grow exponentially in number. Fortunately, we claim that our $K^{\pm}$ components might be found by polynomial $O(m \cdot n \cdot \log_2 n)$ algorithm, as shown in the cited literature. Finally, we can restructure the entire procedure by extracting a component $K_1^{\pm}$ first, before removing it from the original table W and repeating the extraction procedure on the remaining content, thus obtaining components $K_2^{\pm}$, $K_3^{\pm}$,... etc. From now on, statistical parameters and other table characteristics, which empower $(+/-)$ share, arise from components $K_1^-, K_2^-,...$ and $K_1^+, K_2^+,...$ only, and are available to the analyst for illustration purposes, as depicted in the example below.

A.3 Illustration

In the example, we use a sampling highlighting 383 people's attitudes towards 21 phenomenal questions. Each question requires a response on an ordinal scale, with $1 < 2,...,< 5$, where $1 < 2 < 3$ are positive values at the left end, and $3 < 4 < 5$ are negative values at the right end.[4] Hence, our sampling, depicted as a Boolean table, has 383×105 dimensions. As the tuning parameter $k = 5$ was chosen, we also extracted a set of three positive K_1^+, K_2^+, K_3^+ and negative K_1^-, K_2^-, K_3^- components. The actual values in the title and those shares illustrate our positive (green) and negative (red) (+/–) components.

Some typical sampling questions are given below:

[4] Sampling owner (Scanlife Vitality ApS in Denmark) kindly provided us with a permission to use the data for analysis purposes. We are certainly very grateful for such help.

Survey Data Cleaning

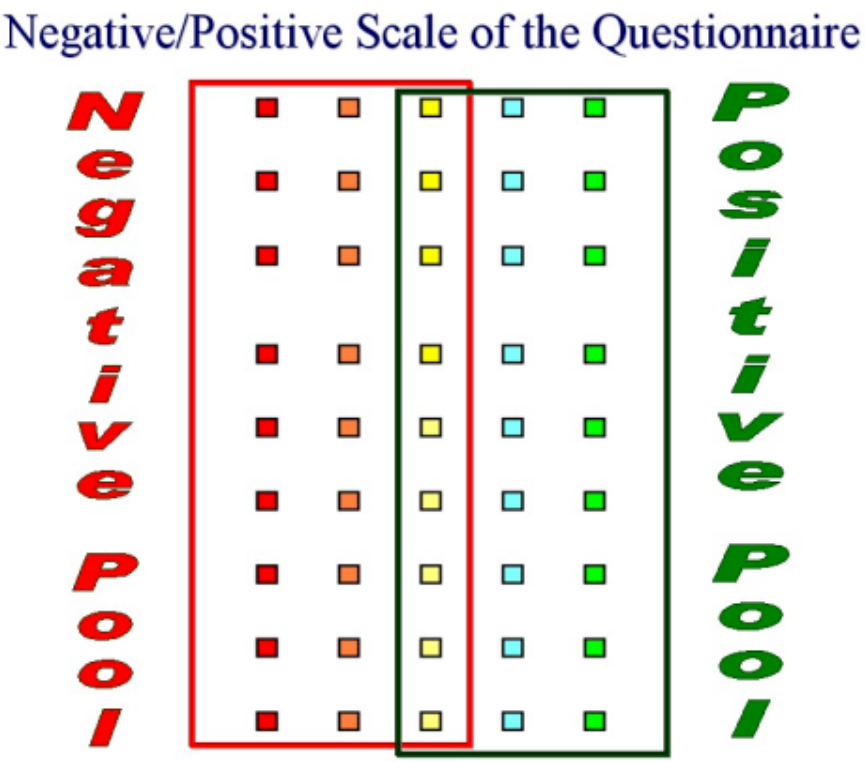
1. Is your behavior slow/quick? – eating, talking, gesticulating,...

   1.1 Absolutely slow
   1.2 Somewhat slow
   1.3 Sometimes slow and sometimes quick
   1.4 Somewhat quick
   1.5 Absolutely quick

2. Are you a person who prefers deadlines/postpones duties?

   2.1 Absolutely always prefer deadlines
   2.2. Often prefer deadlines
   2.3. Sometimes prefer deadlines or sometimes postpone my duties
   2.4. Often postpone my duties
   2.5. Absolutely always postpone my duties

. . . . . .



The figure shows more clearly the methodology of the positive/negative analysis of surveys data tables to identify hidden preferences of respondents. Whatever the analyst is doing to build a negative ordering of the left half of the questionnaire, our negative defining sequence is then compared with similar sequence of the right half of the questionnaire. As a result, two credential scales have been formed, which can then be visualized graphically in two-dimensional coordinate system on the plane.

Acknowledgement

At first glance that being said, our story may seem perhaps frivolous, but we say that it is much easier to suggest something new if the essence of the matter is presented in the form of an allegory, which can be interpreted in such a way as to reveal the hidden meaning of reality.

The fact is that the author of these lines, being a post graduate student of Tallinn University of Technology in 1969-1971, interpreted in his own words (Mullat, 1971a) and invented a general and new procedure of data analysis thanks to a “blind glance” or specific data scoring ideology of his supervisor, prof. Leo Võhandu (Frey and Võhandu, 1966, available online, Accesed: 07 September 2020). Within the framework of this ideology, the author developed a theory that is now known in the literature as "Monotonic Linkage Functions" (Seiarth et al, 2020, available online, Accesed: 22 August 2020), although this model was originally named by the author as "Monotone / Monotonic System".

REFERENCES

Mullat, J.E., a) (1971). ON A MAXIMUM PRINCIPLE FOR CERTAIN FUNCTIONS OF SETS, *in: Notes on Data Processing and Functional Analysis, Proceedings of the Tallinn Polytechnic Institute* (in Russian), Series A, No. 313, Tallinn Polytechnic Institute, pp. 37–44;

b) (1995). A Fast Algorithm for Finding Matching Responses in a Survey Data Table, Mathematical Social Sciences, 30, 195 – 205. http://www.sciencedirect.com/science/article/pii/016548969500780P .

FEJL BETYDER, AT DU FEJLER

EFFEKTIV STYRING AF FORSKNINGSDATA

Du har brug for hurtigt at kunne identificere fejl og mangler i forsknings- og udviklingsprocessen. Som forsker er SAS® Enterprise Guide® dét værktøj, der giver dig grundlaget for at få mest muligt ud af forskningskronerne.

SAS Enterprise Guide giver dig mulighed for avancerede beregninger og analyser. Du kan pege-og-klikke dig frem til et hurtigt overblik samt brugbare resultater, hvilket betyder, at du undgår fejl opstået på grund af ikke valide data.

Læs mere på **www.sas.com/dk/academic**

sas. *The Power to Know.*