Evaluation of the Quality of Survey Data and its Visualization Using Dashboards

Nataliia Komleva
System Software Department
Odesa National Polytechnic University
Odesa, Ukraine
komleva@opu.ua

Vira Liubchenko
System Software Department
Odesa National Polytechnic University
Odesa, Ukraine
lvv@opu.ua

Svitlana Zinovatna
System Software Department
Odesa National Polytechnic University
Odesa, Ukraine
zinovatnaya.svetlana@opu.ua

***Abstract*—Surveys are a popular tool for decision-making support. However, the quality of survey results significantly affects the quality of decisions made on their basis. Firstly based on Shannon's formula, we evaluated the average amount of information received from the group survey. Then we analyzed the information loss caused by omissions and duplication of answers to the questionnaire. Next on the modeled data, we showed the existence of a sufficient respondents' number to reduce the effect of data distortion. Finally, we proposed to use various types of dashboards for visual assessing the appropriateness of data.**

Keywords—survey, questionnaire, data quality, amount of information, dashboard

I. Introduction

Often decision-makers while making decisions need to use the survey results. Surveys are valuable, e.g., for conducting market research, collecting feedback from beta testers, evaluating courses to enhance the teaching quality. The known issue concerning survey using is the quality of the raw data. Poor quality of the raw data leads to the impossibility of decision-making or to the complication, which requires additional computing and human resources, as well as resources of time, memory, etc. In some cases, the wrong decision can be made due to low-quality data. Therefore, the quality of survey data is essential for many areas of human activity. The paper [1] addressed issues of socio-democratic research and states that "accuracy as characteristic of data quality is perhaps the most important issue of all." The review [2] studied public health outcome measurements and highlights that "data, data use, and data collection process, as the three dimensions of data quality, all need to be assessed for overall data quality assessment." Poor data documentation, especially superficial survey reports, also negatively affect the used data quality. The authors of [3] "believe that it is most important to base good research on good data, and good data is distinguished by meaningful methodological documentation."

The decision-makers have to evaluate the survey results to understand whether they are suitable for further use. The data obtained as the survey result can be further processed and used to make strategic decisions based on special rules and specific mathematical methods, such as Analytic Network Process and Linear Programming [4]. An effective decision-making process requires an adequate environment for compensating the existed subjectivities, uncertainty, and inaccuracy [5, 6]. The quality of the data obtained as a survey result is especially important in situations when it comes to determining strategies for achieving long-term goals. Data visualization provides a practical overview of data specific because of well-designed dashboards increase the power of visual information acceptance [7].

The purpose of this paper is to examine information lost due to the distortion of the raw data and the impact of the respondents' number on the survey results that can affect the quality of the data used by decision-makers. The paper also analyses the dashboards' potential for data quality assessment.

In the paper, we consider some properties of survey results, including distortion in raw data, and evaluate the data quality based on the amount of available information. We restrict our consideration only to the answers given on the Likert scale [8].

II. Analysis of the Amount of Information in the Survey Result

Let there is a questionnaire with the set of questions Q, $K=|Q|$. The answers given on the Likert scale can be marked as 1, ..., m, where m is the number of possible ranges. In this case, 1 corresponds to the worst value, and m corresponds to the best value. The average amount of information received per answer corresponds to its entropy and is estimated by the well-known Shannon's formula

$$I_{av}^{1} = H = -\sum_{i=1}^{m}\left(p_i \log_2 p_i\right) \tag{1}$$

where p_i is the probability of the ith answer. In a particular case, the answer options can be considered equally probable with p_i=1/m. Then the amount of information can be estimated using Hartley's measure as $I = \log_2 m$. In the general case, taking into account the background and current survey conditions, we should consider situation $p_i \neq p_j$, where $i \neq j$.

Let there is a representative sample of survey results from R independent respondents on the general population of questionnaires containing K questions. Consider the questions in the questionnaire independent ones. The average amount of information received from the group survey with taking into account the additivity property is calculated as

$$I_{av} = -R \times K \times \sum_{i=1}^{m}\left(p_i \log_2 p_i\right) \tag{2}$$

Thus the average amount of information is directly proportional to the number of respondents and the size of the questionnaire.

III. Distortions in the Raw Data

Often, for both paper and electronic questionnaires, situations of the raw data distortion arise. Technical or personal reasons can cause distortions. In any case, distortions decrease the quality of the raw data. In the paper, we consider typical data distortions and calculate the amount of information for each type of distortion.

Data omission takes place while, for some reason, the answer to some question is not specified. An electronic questionnaire involves technical means that do not allow the transition to the next question without an answer to the current question. However, with a paper questionnaire, the likelihood of such a distortion increases significantly.

Let N is the number of questionnaires that R respondents filled out, $n_1, ..., n_m$ are the quantities of answers with corresponding markers to some question q_k. Let us estimate the average amount of information received per response, under omission one response with an arbitrary marker j.

$$\begin{aligned} I^{omis} &= \frac{n_1 I_1 + n_2 I_2 + \ldots + n_m I_m}{N-1} = \\ &= \frac{1}{N-1}\left(n_1 \left(-\log_2 \frac{n_1}{N-1} \right) + \ldots + \right. \\ &\left. + \left(n_j - 1\right)\left(-\log_2 \frac{n_j - 1}{N-1} \right) + \ldots + n_m \left(-\log_2 \frac{n_m}{N-1} \right) \right) = \\ &= \frac{1}{N-1}\left(n_1 \left(-\log_2 p_1^{omis} \right) + \ldots + n_j \left(-\log_2 p_j^{omis} \right) + \right. \\ &\left. + \ldots + n_m \left(-\log_2 p_m^{omis} \right) - 1\left(-\log_2 \frac{n_j - 1}{N-1} \right) \right) = \\ &= -\sum\nolimits_{i=1}^{m} p_i^{omis} \log_2 p_i^{omis} + \frac{1}{N-1} \log_2 \frac{n_j - 1}{N-1} \end{aligned} \quad (3)$$

Taking into account the fact that the probabilities of sample answers reflect the distribution of the general population, we can assume that $p_i = p_i^{omis}, \forall i = 1, \ldots, m$.

Then

$$\begin{aligned} I^{omis} &= -\sum\nolimits_{i=1}^{m} p_i \log_2 p_i + \frac{1}{N-1} \log_2 \left(\frac{n_j - 1}{N-1} \right) = \\ &= I + \frac{\log_2 p_j}{N-1} = I - \Delta I^{omis} \end{aligned} \quad (4)$$

Because we do not know which answer option is missed, we consider the worst-case:

$$\Delta I^{omis} = \frac{1}{N-1} \max_{0 \le j \le m} \left(\log_2 \frac{n_j}{N-1} \right) \quad (5)$$

The obtained value ΔI^{omis} represents the information loss in the absence of a single answer. Accordingly, for S missed answers, $\Delta I_S^{omis} = S \times \Delta I^{omis}$.

Answers to questionnaire questions may contain irrelevant or uncleaned data. While irrelevant data may be valuable for another task, uncleaned data are not valuable at all. Both cases could be brought to omissions.

While receiving the answers, the mismatch between the time of giving the answer and the time of receiving the answer can be discovered. There can be two types of mistiming: delay when the respondent is not ready to answer at the time of the survey, and technical data mistiming. Also, data that require preliminary processing, e.g., data normalization or data conversion to the required type, can be obtained as the answer. Even if such processing is successful, it takes some time. Thus, if additional resources to correct the distortion are absent, all of the above cases can also be brought to omissions.

Data duplication can appear when generalizing the survey data. Duplication means that the elements of the set R are not unique because the questionnaire of one respondent is counted twice, or one respondent filled out a questionnaire more than once. Let us consider the case of a single duplication of the answer. In this case, the answer with an arbitrary marker j becomes irrelevant; the duplicate should be equated with omission and excluded from consideration. Therefore, the number of useful answers with marker j is n_j, and the total number of answers for which resources have been expended is increased by 1. Then the average amount of information received per one answer is

$$\begin{aligned} I^{dupl} &= \frac{n_1 I_1 + n_2 I_2 + \ldots + n_j I_j + \ldots + n_m I_m}{N+1} = \\ &= -\sum\nolimits_{i=1}^{m} p_i^{dupl} \log_2 p_i^{dupl} \end{aligned} \quad (6)$$

Taking into account the properties of a representative sample, we can assume $p_i = p_i^{dupl}, \forall i = 1, \ldots, m$. The elimination of unnecessary duplicates allows maintaining the quality of information:

$$I^{dupl} = -\sum_{i=1}^{m} \left(p_i \log_2 p_i \right) = I \quad (7)$$

Thus, a priori information about the probabilities of answers, which can be obtained based on conducted questionnaire sessions or expert assumptions, allows evaluating the amount of information lost due to distortion of the raw data.

IV. IMPACT OF THE NUMBER OF RESPONDENTS

Because survey results are always generalized, increasing the number of respondents can improve the data quality. Let m_1, m_2 be the average values for answers to a particular question received from n and n+1 respondents, respectively. The change in the average value caused by adding one respondent is calculated as

$$\begin{aligned} m_2 - m_1 &= \frac{1}{n+1} \sum\nolimits_{i=1}^{n+1} x_i - \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i = \\ &= \frac{1}{n+1} x_{n+1} - \frac{1}{n\left(n+1\right)} \sum\nolimits_{i=1}^{n} x_i \end{aligned} \quad (8)$$

The influence of one answer decreases with an increase in the respondents' number. It tends to zero as n tends to infinity. We can assume, there always exists satisfied sufficient respondents' number n_{lim} that the excess of which will not significantly affect the reliability of the result.

Let us determine the number n_{lim} analytically. Suppose that respondents gave answers on the Likert scale, that is, a set of possible ratings {1,2,3,4,5}. We consider a permissible error of 5% of the minimum estimate, i.e.

$$\frac{1}{n+1}x_{n+1}-\frac{1}{n\left(n+1\right)}\sum_{i=1}^{n}x_i<0,05 \tag{9}$$

Consider a situation when an additional answer differs as much as possible from the other ones, e.g., $x_1=\ldots=x_n=1$, $x_{n+1}=5$. In this case, the effect of the additional answer becomes insignificant for $n>79$.

The analytical definition of n_{lim} was made under acute constrain. Consider the effect of the increase in the respondents' number on the result using model data. The experiment was performed using the add-in Analysis ToolPak for Microsoft Excel 2013. There were generated and examined the sequences of random data under symmetric, single-mode asymmetric, and bimodal discrete distributions. For each distribution, there were generated 50 sequences of random integer numbers in the range from 1 to 5, simulating the answer to the question of the questionnaire. For each sequence, we considered two dependencies – average value as a function of the respondents' numbers and increment of the average value as a function of the respondents' numbers. Both studied dependencies showed similar behavior on each sequence of a particular distribution. Therefore, consideration one sequence of 50 ones under particular distribution will not lead to a loss of generality. Fig. 1 shows the frequency distributions of the sequences, which are discussed below. Each generated sequence contains 50 random numbers.

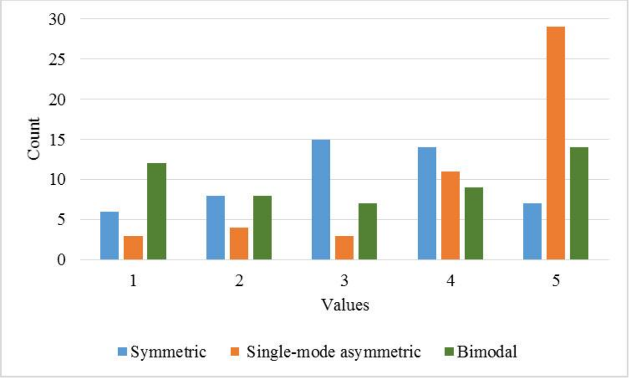


Fig. 1. Distribution histogram for analyzed sequences

Let us consider the dependence of the average value on the number of answers taken into account:

$$m\left(n\right)=\frac{1}{n}\sum_{i=1}^{n}x_i\ . \tag{10}$$

The change in the average values for each of the three sequences is shown in Fig. 2. Each average value $m(n)$ converges to some value with an increase in the number of answers. The tendency does not depend on the distribution of random variables.

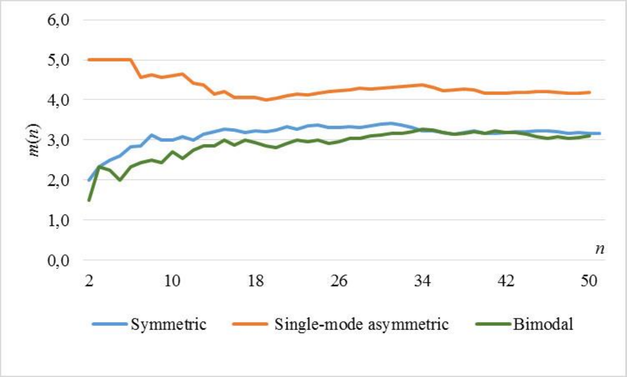


Fig. 2. The changes in average values

Now we will consider how the increment of the average value changes with the increasing of the number of answers; for this, we introduce the function

$$\Delta(n) = m(n) – m(n–1) \tag{11}$$

The changes in increments $\Delta(n)$ for each of the three sequences are shown in Fig. 3.

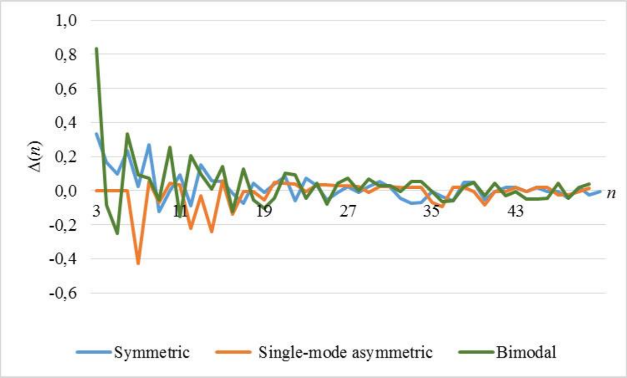


Fig. 3. The changes in increments of average values

Fig. 3 demonstrates that with an increase in the number of answers over 25, the increment of the average value is within the statistical error.

Thus, we can conclude that there always is a sufficient number of answers n_{lim}, the excess of which will not significantly affect the reliability of the result. Therefore, providing n_{lim} answers neutralizes the influence of possible distortions in the source data.

V. Visualization for Assessing the Data Quality

When getting survey results, decision-makers require some form of probing for assessing the appropriateness of data. One approach at assessing the data quality is providing summary visualizations [9] to get a sense of the data distribution and anomalies. Traditionally, various information panels, or dashboards, are used to provide the visual presentation of data grouped by one or more characteristics. Dashboards contain diagrams, explanatory statements, digital symbols, or other elements of infographics.

The main characteristics for dashboards description are [10]:

- quantitative range of signs of different types, acceptable for perception;
- ability to build a functional dependence on one or more arguments;
- set of supported data types for display (numbers, text, video, audio), etc.

Let us look at some examples demonstrated how to use dashboards to assess data quality easily and quickly.

The quality of a survey result with respect to a particular distortion *d* is quantified as the inverted ratio of the number of determined crippled results N_d to the total results count *N*:

$$Q_d = 1 - \frac{N_d}{N} \qquad (12)$$

Therefore, decision-makers should take into account as many quality ratios as many types of distortions appeared. Also, depending on the type of distortion, the data can be classified as useful for decision-making, useless, and potentially useful but needed additional processing. The classification causes the hierarchy of quality ratios.

Instead, a pie chart presents the same information in one picture (Fig. 4), which is more comfortable for humans.

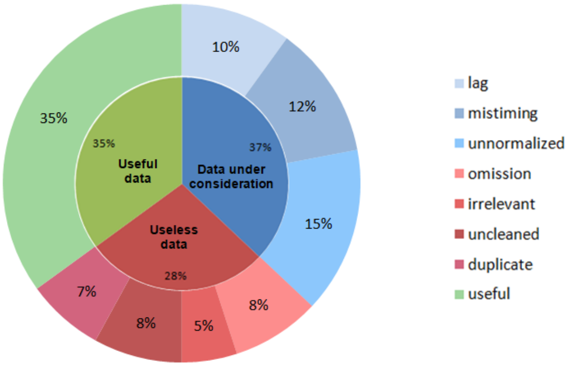


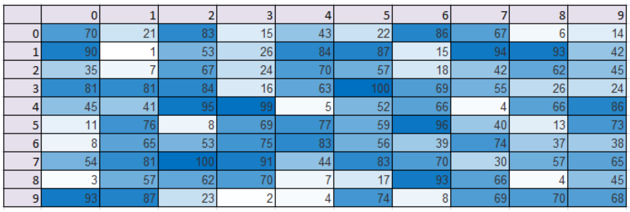
Fig. 4. Detailing the quality of the survey result with a pie chart

Sometimes a visual overview of data quality should be provided. It should specifically convey proportional information on potential errors detected in the answers on particular questions. In such a case, it is convenient to use a monochrome heat map. In the map, the color intensity corresponds to the quality of a particular answer: the most saturated color corresponds to the highest data quality. The quality level for the particular answer is evaluated as

$$level = \frac{N_{+} - \left(N_{-} + N_{+/-}\right)}{N_{+}} \times 100 \qquad (13)$$

where N_{+}, N_{-}, and $N_{+/-}$ are contributions of useful, doubtful, and useless data, respectively.

For example, Fig. 5 gives a visual overview of survey results. The survey used the questionnaire with 100 questions. Each number in the cell indicates the quality level of the answer to the particular question.



	0	1	2	3	4	5	6	7	8	9
0	70	21	83	15	43	22	86	67	6	14
1	90	1	53	26	84	87	15	94	93	42
2	35	7	67	24	70	57	18	42	62	45
3	81	81	84	16	63	100	69	55	26	24
4	45	41	95	99	5	52	66	4	66	86
5	11	76	8	69	77	59	96	40	13	73
6	8	65	53	75	83	56	39	74	37	38
7	54	81	100	91	44	83	70	30	57	65
8	3	57	62	70	7	17	93	66	4	45
9	93	87	23	2	4	74	8	69	70	68

Fig. 5. Presentation of quality levels of particular answers with a heat map

After the identification and registration of the raw data distortions, decision-maker should minimize their impact by the selection and application of additional data processing or various tactics designed to manage the sources of distortion.

VI. Conclusion

Data quality significantly affects the quality of decisions made on their basis. We showed how distortions in the survey results affect the amount of information. However, a large number of answers can neutralize this impact. Since the survey with a large number of respondents is resource-intensive, a sufficient number of respondents is of interest. Using modeled data, we showed that 25 or more respondents could neutralize possible distortions in the data. Finally, the issue of visualization for assessing the data quality was considered. The visualization simplifies assessing the appropriateness of data for decision-makers.

References

[1] L. Singh, “Accuracy of Web Survey Data: the State of Research on Factual Questions in Surveys,” Information Management and Business Review, vol. 3(2), pp. 48-56, August 2011.

[2] H. Chen, D. Hailey, N. Wang, and P. Yu, “A Review of Data Quality Assessment Methods for Public Health Information Systems,” International Journal of Environmental Research and Public Health, vol. 11(5), pp. 5170–5207, May 2014.

[3] A. Jedinger, O. Watteler, and A. Förster, “Improving the Quality of Survey Data Documentation: A Total Survey Error Perspective,” Data, vol. 3(4), 45, October 2018.

[4] D. Wollmann, and M. T. A. Steiner, “The Strategic Decision-Making as a Complex Adaptive System: A Conceptual Scientific Model,” Complexity, vol. 7, pp. 1–13, January 2017.

[5] A. H. Ahmed, H. Bwisa, R. Otieno, and K. Karanja, “Strategic decision making: process, models, and theories,” Business Management and Strategy, vol. 5(1), pp. 78–104, 2014.

[6] T. L. Saaty, “About a hundred years of creativity in decision making,” International Journal of the Analytic Hierarchy Process, vol. 7(1), March 2015.

[7] S. Few, Information Dashboard Design: Displaying Data for At-a-glance Monitoring. El Dorado Hills, CA: Analytics Press, 2013.

[8] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, “Likert Scale: Explored and Explained,” British Journal of Applied Science & Technology, vol. 7(4), pp. 396-403, 2015.

[9] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, “Profiler: Integrated statistical analysis and visualization for data quality assessment,” in Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI’12). New York, ACM, 2012, pp. 547–554.

[10] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, “What Do We Talk About When We Talk About Dashboards?” IEEE Transactions on Visualization and Computer Graphics, vol. 29(1), pp. 682–692, 2019.