Conditional histogram analysis of discrete questionnaire data

Tetiana Reznychenko, Evženie Uglickich, Ivan Nagy

***Abstract*—The paper deals with the analysis of histograms of discrete data collected in questionnaires obtained for individual realizations of the target variable. The main aim of the analysis is to explore the influence of combinations of explanatory variables, represented by responses to the questionnaire, on the behavior of the target variable of the questionnaire. In this paper, an automated approach to histogram comparison is proposed based on coding combinations of data and detecting significant differences in frequencies using the Marascuilo procedure. This is the main contribution of the paper. The approach is validated using a simulated questionnaire in which respondents answered regarding their intention to purchase an electric vehicle subject to finance, leasing, and charging availability, as well as their driving style. The results of the experiments are demonstrated.**

***Index Terms*—conditional histograms, discrete data, Marascuilo procedure, questionnaire analysis**

I. Introduction

Questionnaires are an effective tool for collecting data in various application areas (for example, accident prediction, voter preferences, customer service, etc.). The questionnaires produce sets of discrete data that should be analyzed.

Data analysis found in this field is of several types. One of them is a part of descriptive statistics using (i) simple univariate analysis both of nominal and ordinal variables including the estimation of proportions, (ii) standard chi-square goodness-of-fit test and graphical visualization (bar, pie charts, histogram) [1]–[4] along with the mean estimation for ordinal data [1], and (iii) the bivariate extension of estimating and testing proportions [1], [3], [4].

Comparative analysis of discrete questionnaires is based on hypothesis testing to investigate the data distribution, search for dependence/independence of variables, etc. The analysis of contingency tables in this field is based on testing associations between the variables under certain assumptions. The methods are divided between those for (i) *two-way tables*, e.g., the chi-square test of independence [1]–[3], [5], McNemar test for binary variables [6]; Fisher's exact test for nominal variables [3]; Goodman and Kruskal's rank correlation measures for ordinal and nominal data [7], etc., and (ii) *multiway tables*, e.g., the Cochran-Mantel-Haenszel test dealing with the estimation of the odds ratio and relative risk [1], [4], simple log-linear models [4], [5], [8], etc.

This paper deals with inductive analysis aimed at modeling the relationship between a target variable and several explanatory variables. In this area, the existing methods can be divided as follows:

- modeling a binary target variable using Generalized Linear Models (GLM), namely, the logistic, probit [1], [2], [5], [9], [10] and gompit regressions [10] directed at the classification of continuous explanatory data and the application of the logistic regression to indicator-based discretized explanatory data [3];
- GLMs for a multinomial target variable:
  - the multinomial logit regression, e.g., [2], [5] intended primarily for *nominal* target variables, but also used for Likert-type scaled or other rate scaled ordinal data [4];
  - the cumulative logit model for *ordinal* target variables [1], [3] (here, the application of a linear regression to ordinal data treated as continuous random variables [1], [11] or multinomial logit regression ignoring the knowledge of ordering the target values can be also met [4], [5], [9]);
  - the Poisson and negative binomial regression models [1], [3], [4] for *count* target variables as well as their zero-inflated versions [12], the use of linear regression techniques [5];
- techniques of coding discrete variables (e.g., dummy, repeated coding, etc.) and using multiple indicators for a linear regression [4], [11];
- the item response theory with the Rasch model elaborated for *binary* target variables and partial credit, rating scale and graded response models using a logit link function for *multinomial* ordinal [13], [14] and nominal [14], [15]

The project was partially supported by the project TAČR FW06010535, the StorAIge project, and the corresponding Czech institutional support project No. 8A21009. This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007321. The JU receives support from the European Union's Horizon 2020 research and innovation program and France, Belgium, Czech Republic, Germany, Italy, Sweden, Switzerland, Turkey.

T. Reznychenko is with the Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 11000 Prague, Czech Republic (e-mail: tetiana.reznychenko@cvut.cz).

E. Uglickich is with the Department of Signal Processing, Institute of Information Theory and Automation of the CAS, Pod vodárenskou věží 4, 18208 Prague, Czech Republic (e-mail: uglickich@utia.cas.cz).

I. Nagy is with both the Faculty of Transportation Sciences, Czech Technical University, 11000 Prague, Czech Republic, and the Department of Signal Processing, Institute of Information Theory and Automation of the CAS, 18208 Prague, Czech Republic (e-mail: nagy@utia.cas.cz).

979-8-3503-2162-3/23/$31.00

variables along with the Mokken scaling analysis [16];
- analysis of ordinal Likert scale variables using structural equation modeling [17];
- Bayesian categorical analysis, see [5], [18]–[20], etc.;
- imputation methods for treating missing data in questionnaires, e.g., [1], [2], [21], [22];
- clustering and classification of discrete data by means of data mining techniques, such as decision tree [23], Bayesian networks [24], neural networks [25], k-nearest neighbors [26], fuzzy rules [27], naive Bayes classifiers [28], [29], genetic algorithms and model-based methods including the use of discrete mixture models such as latent class and Rasch mixture models [5], Poisson and negative binomial mixtures [18], mixtures of Poisson regressions [5], [18], mixtures of logistic regressions for binary data [18], Poisson-gamma and beta-binomial models [5], [18] as well as Dirichlet mixtures [30], [31]. The estimation of the mentioned mixtures is solved primarily using the iterative expectation-maximization (EM) algorithm [32]. Algorithms of recursive estimation of categorical mixtures with conjugate prior Dirichlet distributions are elaborated in [19], [20], [33].

Despite the significant number of research methods of discrete data analysis, this field still needs novel solutions that can be used for modeling specific questionnaire data. Working with a questionnaire with many explanatory variables, where each of them has several possible answers, it is necessary to find typical groups of them (which means to explore their probability functions) and investigate which combinations of explanatory variables lead to certain values of the target variable. The presented paper focuses on this specific task through analysis and comparison of histograms of combinations of explanatory variables obtained for individual realizations of the target variable. In the case of a high number of values of the variables involved (i.e. multiple choice responses), the dimension of the questionnaire table increases and therefore an automated procedure should be developed to accomplish this task.

The layout of the paper is organized as follows. Section II-A formulates problem to be solved. Section II-B introduces the proposed methodology. The results of experiments with simulated data are provided in Section III. Conclusions can be found in Section IV.

II. Theoretical Background

A. Problem Formulation

Let us have a questionnaire, which contains a set of the explanatory data $X = \{[x_1, x_2, \ldots, x_n]_t\}_{t=1}^{T}$ and the target variable $Y = \{y_t\}_{t=1}^{T}$. The explanatory variable X is the n-variate discrete random variable. Each of their realizations $x_{1;t}, x_{2;t}, \ldots, x_{n;t}$ at time $t = 1, 2, \ldots, T$ corresponds to answers to n questions. Each of the questions has the set of their possible realizations $\{1, \ldots, N_{x_i}\}$, $i = 1, 2, \ldots, n$, which is not limited to be of the same dimension, i.e., N_{x_i} can be different.

The target discrete variable Y has a set of possible realizations $y_t \in \{1, 2, \ldots, N_y\}$.

The task to be solved is formulated as follows: *find the combinations of X for which there are significant differences between the values of Y.*

In the context of the questionnaire analysis, this means finding out how the answers differ, how significant the differences are, and what probability functions correspond to each answer.

The main idea of the presented approach along with a simple example is shown below.

B. Conditional Histogram Analysis of Questionnaire Data

In the considered context, the relationship of the explanatory and target variables is described by the joint probability function (pf), which is decomposed according to the chain rule [34]

$$f(Y, X) = f(X|Y)\, f(Y), \tag{1}$$

where $f(X|Y)$ is the conditional pf, which describes the behavior of the explanatory variables depending on the target variable Y, and $f(Y)$ is the marginal pf of Y, which is its histogram.

The main idea is to analyze the differences in the conditional pfs $f(X|Y)$ existing for each value of Y. The sets of values of all variables in X are large, which leads to a table of a huge dimension. That is why it is necessary to code the combinations of individual values from X to the auxiliary coding variable g, i.e.,

$$f(X|Y) = \begin{cases} f(X|Y=1) \rightarrow f(g|Y=1) \\ f(X|Y=2) \rightarrow f(g|Y=2), \\ \ldots \\ f(X|Y=N_y) \rightarrow f(g|Y=N_y). \end{cases} \tag{2}$$

where $g = [g_1, g_2, \ldots, g_{N_g}]$, $N_g = \prod_{i=1}^{n} N_{x_i}$. All of the conditional pfs $f(g|Y=j)$, $j = 1, 2, \ldots, N_y$ in the form of histograms should be compared.

If the differences in frequencies for a given value of g (representing a combination of individual variables from X) are not significant, the result does not bring any new knowledge about the behavior of Y. However, in the case of significant differences, it indicates that the given combination of X influences the variable Y. From a practical point of view, it is important which respondents create the group of these combinations and how it is possible to change the values of Y via the combinations of X.

The following simple example illustrates the presented idea. Let the values of Y at time t be $y_t \in \{1, 2\}$, and $x_{1;t} \in \{1, 2, 3\}$ and $x_{2;t} \in \{1, 2\}$. The entire explanatory data set X is divided into two groups according to the realizations of Y, where each row is coded by the value of the new coding variable g:

$$Y = 1$$

$x_{1;t}$	1	1	2	2	3	3
$x_{2;t}$	1	2	1	2	1	2
g	1	2	3	4	5	6
$h(g)$	5	7	**51**	8	3	4

$Y = 2$

$x_{1;t}$	1	1	2	2	3	3
$x_{2;t}$	1	2	1	2	1	2
g	1	2	3	4	5	6
$h(g)$	2	4	3	5	**98**	6

and where $h(g)$ are histograms of g (i.e., individual combinations of $x_{1;t}$, $x_{2;t}$).

Here, significant frequencies correspond to $g = 3$ (for $Y = 1$) and $g = 5$ (for $Y = 2$). It means that $g = 3$ composed from the combination of $x_{1;t} = 2,\ x_{2;t} = 1$ induce the value $y_t = 1$ and similarly, $g = 5$ corresponding to $x_{1;t} = 3,\ x_{2;t} = 1$ leads to $y_t = 2$. The rest of the observed frequencies are insignificant.

So, if $y_t = 1$ is positive and $y_t = 2$ negative feature of the target variable Y (for instance, the accident occurrence), from a practical point of view, it is necessary to replace $x_{1;t} = 3,\ x_{2;t} = 1$ by $x_{1;t} = 2,\ x_{2;t} = 1$ (which leads to replacing $x_{1;t} = 3$ by $x_{1;t} = 2$). For example, if one of these variables is the number of lanes and the second one is the speed limit, the replacements to be needed are to change (increase) the number of lanes and lower the speed limit, etc. Another alternative is to search for values of g (combinations of X), which are responsible for a certain number of serious traffic accidents, and investigate them.

In this paper, the automated approach of the histogram comparison is based on coding the combinations of values and finding different frequencies with the help of the Marascuilo procedure [35]. The individual steps of the procedure are summarized as follows:

1) Classify the explanatory data X among the values of the target variable Y.
2) Code each row of combinations of the values X as the value of the coding variable $g \in \{1, 2, \ldots, n\}$.
3) Obtain the histograms of g for each Y.
4) Apply the Marascuilo procedure to test the differences of columns in individual histograms of g.
5) Decode the variable g back to the original values of X to investigate which combinations cause the significant differences.

The final step for selected combinations must be done manually. However, the number of the found combinations will not be so large.

The presented approach of the histogram analysis was tested in the open source software for numerical computations Scilab (see www.scilab.org).

C. Marascuilo procedure

Marascuilo procedure can be used for the comparison of equality of several proportions $p = [p_1, p_2, \ldots, p_n]$. It gives results for all individual couples p_i, p_j.

The statistics are the absolute value of differences

$$s_{i,j} = |p_i - p_j|, \forall\, i, j \in \{1, 2, \ldots, n\}, i \neq j. \tag{3}$$

The critical values are

$$r_{i,j} = \sqrt{\chi^2_{1-\alpha,n-1}} \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}, \tag{4}$$

where χ^2 is critical value of the χ^2-distribution, α is the confidence level, n is the number of proportions and n_i, n_j are numbers of data from which the proportions p_i, p_j have been computed.

The result is

$$M_{i,j} = \begin{cases} 1 & \text{for } s_{i,j} \geq r_{i,j} \\ 0 & \text{for } s_{i,j} < r_{i,j}. \end{cases} \tag{5}$$

The proportions p_i, p_j for which $M_{i,j} = 1$ are different [36].

III. Experiments

The aim of the experiments was to validate the approach by finding the differences in conditional histograms in simulations in the correct combinations of explanatory data. The simulated set of discrete questionnaire data concerning an intention of a potential customer to purchase an electric vehicle was prepared for the experiments.

Two configurations of the questionnaire were used for the validation. In the first case, the target variable was a customer's intention to buy an electric vehicle (EV) $\in \{1, 2\}$ with 1 denoting the intention to buy and 2 not to buy EV.

For this configuration, we have used two models with two columns of switching probabilities to express the intention of a customer to buy or not to buy EV. With the help of these models, we used the random generator in Scilab to simulate the values of the target and explanatory variables.

The explanatory variables were as follows:

- finances $\in \{1, 2\}$, where 1 means that customers have sufficient financial means to buy EV and 2 – they do not have the financial capacity,
- credit $\in \{1, 2\}$, where 1 – they agree to ask, and 2 – customers do not agree to ask a bank for credit or leasing,
- charging $\in \{1, 2\}$, where 1 – they have a possibility, and 2 – customers do not have a possibility to charge EV close to home or workplace,
- driving style of the customers $\in \{1, 2\}$, where 1 – shorter distances only in a city, and 2 – mostly longer distances out of a city.

In the second case, three possible realizations were considered: 1 – customers want to buy EV, 2 – they do not want to buy EV, 3 – not sure about buying EV. In this case, three models of switching probabilities were used for the simulation. The explanatory variables were defined in the same way as in previous case. The number of data generated will be mentioned later in the description of individual experiments.

The results of validating the approach obtained using both configurations are presented below.

A. Experiments with Two Values of a Customer's Intention of Buying EV

Here, the important value of the target value is $Y = 2$, which means that customers are more likely not to buy EV. The analysis of the combinations of respondents' answers leading to this value can be useful from a practical point of view and provide an understanding of which changes should be made.

For this part of the experimental part of the work, two options were considered: (i) with 5.000 simulated values, and (ii) with 10.000 values, which correspond to the number of people that completed the questionnaire. The histograms corresponding to 5.000 respondents and to intention of customers to purchase and not to purchase EV can be seen in Fig. 1 (top and bottom).

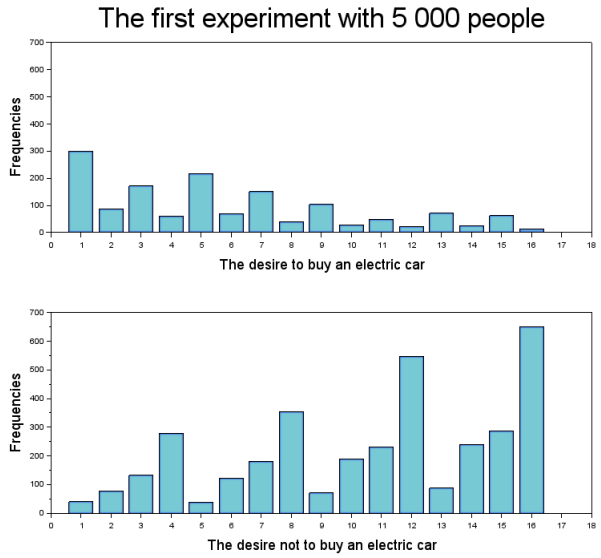

Fig. 1. Histograms for two values of the target variable and 5.000 participants

The frequencies different in values of the coding variable g were obtained as follows:

$$1 \;\; 4 \;\; 5 \;\; 8 \;\; 10 \;\; 11 \;\; 12 \;\; 14 \;\; 15 \;\; 16.$$

The approach found the following combinations of the explanatory data, corresponding to them:

1	1	1	1	1
4	1	1	2	2
5	1	2	1	1
8	1	2	2	2
10	2	1	1	2
11	2	1	2	1
12	2	1	2	2
14	2	2	1	2
15	2	2	2	1
16	2	2	2	2

Comparing the top and bottom histograms, for instance, for the value of $g = 16$, it can be seen that the frequencies for the intention to buy and not to buy EV are highly different: about 30 vs 650. The algorithm evaluated this combination as influencing the target variable. Unlike $g = 16$, for instance, the algorithm did not select the value of $g = 3$, because in the histograms the difference in frequencies between $Y = 1$ and $Y = 2$ is less than 100. It means that this combination of explanatory variables is not significantly different.

Similarly, the testing of the proposed approach with 10.000 values has provided the same correct results expected from known simulations, see Fig. 2.

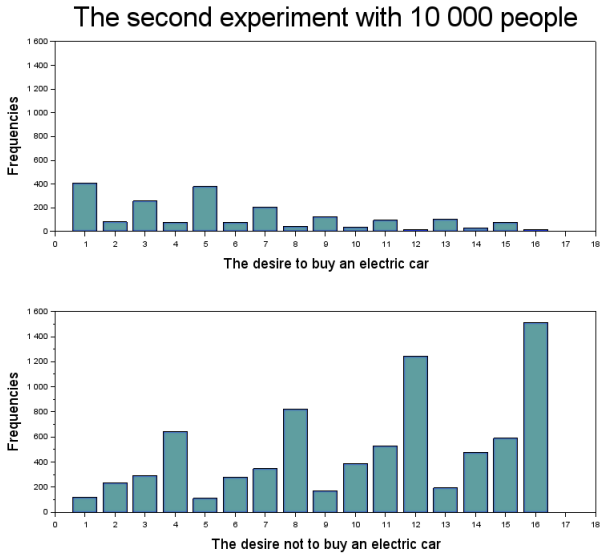

Fig. 2. Histograms for two values of the target variable and 10.000 participants

This means that for this configuration of the simulated data set, the conditional histogram method works similarly for 5.000 and 10.000 values.

B. Experiments with Three Values of a Customer's Intention to Buy EV

In this part, the value $Y = 2$ is still the subject of interest, however, in addition, respondents have a choice to answer that they are not sure about buying EV ($Y = 3$).

Here, three options were considered, where the number of simulated respondents was 3.000, 5.000, and 10.000. Here, three histograms were compared, and the combinations are different. The results of the conditional histogram comparison for 3.000 values are given in Fig. 3. In this case, the frequencies different in values of g were

$$5 \;\; 12 \;\; 16$$

with the following combinations of explanatory data:

5	1	2	1	1
12	2	1	2	2
16	2	2	2	2

For 5.000 participants, the histograms compared can be found in Fig. 4. Here, the values of g were

$$1 \;\; 5 \;\; 7 \;\; 8 \;\; 12 \;\; 15 \;\; 16$$

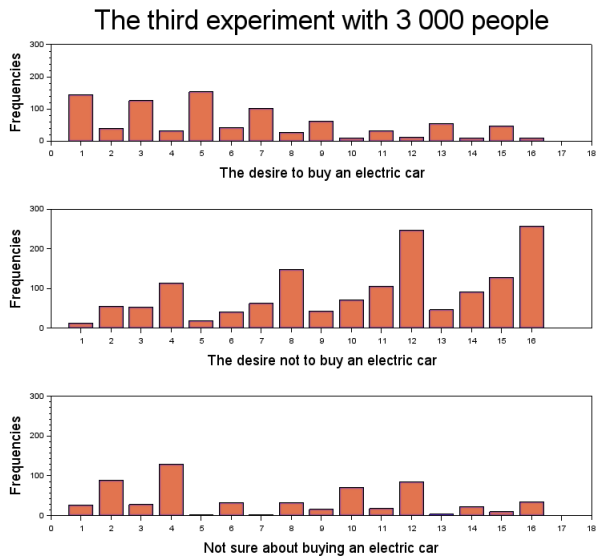

Fig. 3. Histograms for three values of the target variable and 3.000 people.

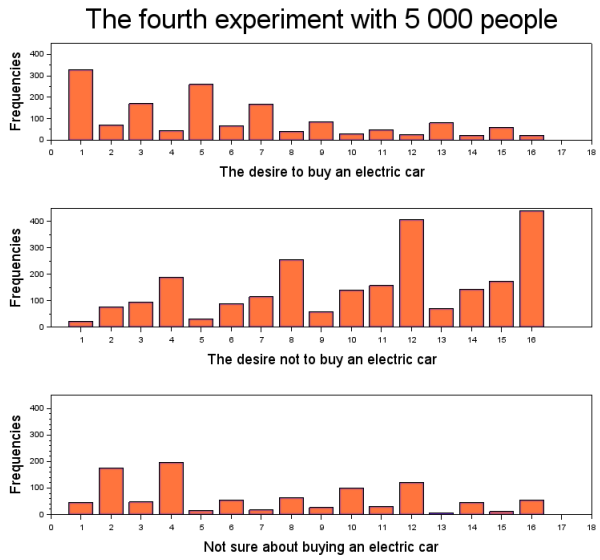

Fig. 4. Histograms for three values of the target variable and 5.000 people.

and the combinations of the explanatory data were

1	1	1	1	1
5	1	2	1	1
7	1	2	2	1
8	1	2	2	2
12	2	1	2	2
15	2	2	2	1
16	2	2	2	2

During the validation with 10.000 people shown in Fig. 5, the values of g found by the approach were

$$1 \;\; 5 \;\; 7 \;\; 8 \;\; 11 \;\; 12 \;\; 14 \;\; 15 \;\; 16$$

and the combinations of data corresponding to them were

1	1	1	1	1
5	1	2	1	1
7	1	2	2	1
8	1	2	2	2
11	2	1	2	1
12	2	1	2	2
14	2	2	1	2
15	2	2	2	1
16	2	2	2	2

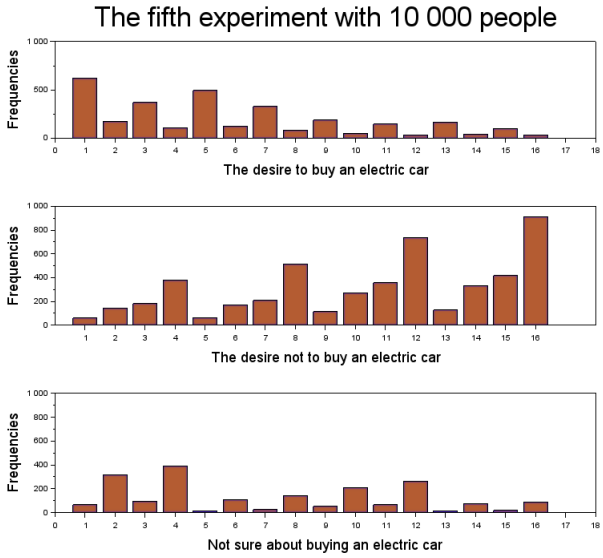

Fig. 5. Histograms for three values of the target variable and 10.000 people.

The histograms, which are shown in Fig. 3, Fig. 4, and Fig. 5, are similar. However, the method identified nine combinations that are different for 10,000 people. We got three combinations with 3,000 people in Fig. 3, and seven combinations with 5,000 people in Fig. 4. In Fig. 5, the frequencies are different for g = 3 and g = 7. These combinations may have an impact on the target variable, but the algorithm did not evaluate them as significant.

The significant frequencies for the experiment correspond to $g \in \{1, 5\}$ for $Y = 1$, and $g \in \{12, 16\}$ for $Y = 2$. This means that g is composed of the combinations

1	1	1	1	1
5	1	2	1	1
12	2	1	2	2
16	2	2	2	2

If $Y = 2$ is the feature of the interest, because the factors influencing intentions of the customers not to purchase EV have to be explored, it would be necessary to replace $x_{1;t} = 2,\ x_{3;t} = 2,\ x_{4;t} = 2$ by $x_{1;t} = 1,\ x_{3;t} = 1,\ x_{4;t} = 1$. From a practical point of view, it describes the reasons that restrict customers from buying EV. As it was expected from the simulated data set, the changes that should be made to convince customers are (i) increasing their purchasing power, (ii) increasing the number of EV charging stations, and (iii)

adapting their driving style to shorter distances in the city, which may suggest that they should have an EV more as a second car in the family. The bank credit $x_{2;t}$ is not a significant explanatory variable in this case.

C. Discussion

The focus of the paper was to test the proposed approach using simulated data. The goal was successfully achieved. The results obtained from the experiments performed show that the method has been successfully applied to the analysis of conditional histograms. The task is very relevant in practical areas, for example, in transportation sciences. Potential applications of the method can be seen in the optimization of transport routes, or the reduction of carbon vehicles that lead to air pollution, etc. In the next paper, we plan to test the approach using real data.

IV. Conclusion

The presented study focused on the comparison of frequencies in histograms of explanatory data obtained for the case of different values of the questionnaire target variable. The aim of the study was to recognize the significant differences in histograms and to find the combinations of explanatory data, having an impact on the target variable. The obtained results show that this algorithm works more accurately on a sample of 3.000 and 5.000 people and three values of the target variable. For the approach, it is important to define a criterion by which it is possible to say that the values are really different. The Marascuilo procedure showed adequate results of the histogram comparisons, which means that the goals of the study were achieved.

The main contribution of the paper is the automated approach of the histogram comparison based on coding the combinations of values and finding different frequencies with the help of the Marascuilo procedure.

The question of what can be considered as different frequencies in the histograms is very sensitive. For the time being, a statistical approach based on the Marascuilo test was used. However, the essence of the problem is: when can groups of answers be considered different. A clear formulation for this question and an approach to its solution is needed.

References

[1] S.G. Heeringa, B.T. West, P.A. Berglung. "Applied Survey Data Analysis". Chapman & Hall/CRC, 2010.

[2] W. Tang, H. He, X. M. Tu. "Applied Categorical and Count Data Analysis". Chapman and Hall/CRC, 2012.

[3] A. Agresti. "An Introduction to Categorical Data Analysis". 3rd Ed. Wiley, 2018.

[4] B. Falissard. "Analysis of Questionnaire Data with R". Chapman & Hall/CRC, Boca Raton, 2012.

[5] A. Agresti. "Categorical Data Analysis". 3rd Ed. John Wiley & Sons, 2012.

[6] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 1947, 12(2): 153-157.

[7] L. A. Goodman, W. H. Kruskal. Measures of association for cross classifications III: approximate sampling theory. Journal of the American Statistical Association, 1963, 58(302): 310-364.

[8] M. E. Stokes, C.S. Davis, G. G. Koch. "Categorical Data Analysis Using SAS". 3rd Ed. SAS Institute, 2012.

[9] D. W. Hosmer, S. Lemeshow. "Applied Logistic Regression". 2nd Ed. Wiley-Interscience, 2000.

[10] P. D. Allison. "Logistic Regression Using SAS: Theory and Application". 2nd Ed. SAS Institute, 2012.

[11] T.Z. Keith. "Multiple Regression and Beyond. An Introduction to Multiple Regression and Structural Equation Modeling". 3rd Ed. Routledge, New York and London, 2019.

[12] J. S. Long, J. Freese. "Regression Models for Categorical Dependent Variables Using Stata". 3rd Ed. Stata Press, 2014.

[13] F. Bartolucci, S. Bacci, M. Gnaldi. "Statistical Analysis of Questionnaires: A Unified Approach Based on R and Stata". Chapman & Hall/CRC, Boca Raton, 2016.

[14] M. L. Nering, R. Ostini. "Handbook of Polytomous Item Response Theory Models". Routledge, 2010.

[15] R. D. Bock. "The nominal categories model". In: W. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory. New York: Springer, 1997, pp. 33-50.

[16] K. Sijtsma, L.A. van der Ark. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. British Journal of Mathematical and Statistical Psychology, 2017, 70, 137-158.

[17] Z. Awang, A. Afthanorhan, M. Mamat. The Likert scale analysis using parametric based Structural Equation Modeling (SEM). Computational Methods in Social Sciences. 2016, vol.4, 13-21.

[18] P. Congdon. "Bayesian Models for Categorical Data". John Wiley & Sons, 2005.

[19] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. "Optimized Bayesian Dynamic Advising: Theory and Algorithms". Springer, London, 2006.

[20] I. Nagy, E. Suzdaleva. "Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components", SpringerBriefs in Statistics. Springer International Publishing, 2017.

[21] M.R. Stavseth, T. Clausen, J. Røislien. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. SAGE Open Med. 2019 Jan 8; 7:2050312118822912.

[22] R. J. A. Little, D. B. Rubin. "Statistical Analysis with Missing Data". 3rd Ed. Wiley, 2019.

[23] S.L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1994. Mach Learn 16, 235-240.

[24] R. E. Neapolitan. "Learning Bayesian Networks". Pearson, 2019.

[25] C. C. Aggarwal. "Neural Networks and Deep Learning: A Textbook". Springer, 2018.

[26] T. M. Cover and P. E. Hart, Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13(1): 21-27.

[27] M. R. Berthold, B. Wiswedel, T. R. Gabriel. Fuzzy logic in KNIME – modules for approximate reasoning. International Journal of Computational Intelligence Systems, 2013, 6(1): 34-45.

[28] D. Forsyth. "Applied Machine Learning". Springer, 2019.

[29] S. M., Basha, D. S. Rajput, R. K. Poluru, S. B. Bhushan, S. A. K. Basha. Evaluating the performance of supervised classification models: decision tree and naïve Bayes using KNIME. International Journal of Engineering & Technology, 2018, 7(4.5): 248-253.

[30] N. Bouguila, W. Elguebaly. Discrete data clustering using finite mixture models. Pattern Recognition, 2009, 42(1): 33-42.

[31] Y. Li, E. Schofield, M. Gönen, A tutorial on Dirichlet process mixture modeling. Journal of Mathematical Psychology, 2019, 91: 128-144.

[32] M. R. Gupta, Y. Chen. Theory and Use of the EM Method. (Foundations and Trends(r) in Signal Processing). Now Publishers Inc., 2011.

[33] M. Kárný. Recursive estimation of high-order Markov chains: Approximation by finite mixtures. Information Sciences, 2016, 326: 188-201.

[34] V. Peterka. "Trends and Progress in System Identification", in Bayesian system identification. Oxford: Pergamon Press, 1981, pp. 239–304.

[35] D. Adeidia and S. Appiah. Assessment of Hypertension-Induced Deaths in Ghana: A Nation-Wide Study from 2012 to 2016. Journal of Data Analysis and Information Processing, 2020, 8, 158-170.

[36] S. T Wagh, N. A. Razvi. Marascuilo Method of Multiple Comparisons (An Analytical Study of Caesarean Section Delivery). International Journal of Contemporary Medical Research, 2016, 3(4): 1137-1140.