UDC 004.042

A. Stenin, V. Pasko, E. Melkumian, S. Stenin

RESTORE TABLES WITH PARTIALLY MISSING DATA

Annotation. An approach to solving the problem of converting a table with incomplete source data or containing values that do not correspond to the expected measurement result into a representative sample considered. The existing methods of missing data recovery are analyzed, the method of homogeneous groups for missing data recovery in the original binary data table presented.

Keywords: method of questionnaire, questionnaire table of binary data, homogeneous data groups, the method of homogeneous groups, restoration of missing data

Introduction

The problem of processing data gaps has to be faced in a wide variety of statistical analysis applications [1,4,5]. This is also typical for the analysis using method of questionnaire [2,3]. When filling out the questionnaire, the expert may not answer all questions for various reasons (missed, forgot, did not answer, did not know the answer, etc.). If one question not answered by the majority of experts, it is a lack of questionnaire and you need to adjust this question and repeat the survey. If only a few experts missed this question, you can try to determine the most likely answers in this situation. This is to ensure that all data is involved in the processing. Interviewers tend to seek to eliminate gaps as quickly as possible in order to subsequently process «complete» data by standard means, with little thought that such an approach can lead to a strong difference in statistical findings when there are gaps in the data and when there are no gaps. The most common methods of analyzing data with omissions are the exclusion of incompetent observations (i.e., containing at least one omission) and traditional methods of filling in omissions. These methods generally have low efficiency, lead, as a rule, to bias and insolvency, to violation of the significance levels of the criteria and other distortions of statistical conclusions, do not have resistance to the distribution of omissions.

Very often these data is formed in the form of rectangular tables. Rows (columns) of the data table correspond to different sources of information, depending on the purpose. Then the columns (rows) are the variables under study (traits, scores, ratings, etc.). The elements of the table are real numbers (numerical characteristics of products and goods), binary numbers (1,0) or (+, -), discrete numbers (for example, when ranking quality evaluation criteria) etc.

 ISSN 1560-8956

Overview of existing solutions

This article devoted to the analysis and recovery of binary data in cases where the table of the values of variables is missing. They can absent due to the fault of the expert, as mentioned above, for technical reasons due to equipment failure, as well as, when the opinion of some sources of information not preference for one criterion over another. There are four main groups of data processing methods with omissions [6,7].

This article devoted to the analysis and recovery of binary data in cases where the table of the values of variables is missing. They can absent due to the fault of the expert, as mentioned above, for technical reasons due to equipment failure, as well as, when the opinion of some sources of information not preference for one criterion over another. There are four main groups of data processing methods with omissions [6,7].

The methods of exclusion. In the absence of some variables, they removed from the population and the remaining data is processed. These methods are easy to implement and can be satisfactory with a small number of omissions and a large population of data. However, sometimes lead to large displacements of estimates and are not always effective. In the most software packages for statistical analysis allowed the selection of the missing elements in the data table by using the specific code. To highlight missing elements of various kinds, you may need several codes, such as «don't know», «refuse to answer», «invalid response». The following methods used.

A method for the analysis of complete observations.

Processing of data of complete observations reduced to the use of only those observations in which all variables are present. The advantages of this method are its simplicity (because you can directly apply standard methods of analysis for complete data) and comparability of one-dimensional statistics (because they are all calculated by one set of observations). The disadvantage of this approach is the loss of information, excluding incomplete observations, and the decrease in data can be significant (especially for small samples).

A method of available observations.

All available values are used. A natural generalization of the method of available observations in the multidimensional case is the pair methods of available observations, when the measure of dependence between the variables *Ej* and *Ek* calculated from the observations *i* for which both *eij* and *eik* are present. The disadvantage of this method is that the set of observations on which the sample based varies from feature to feature in accordance with the omission matrix.

Methods of filling. The omissions filled in and the «*full*» data processed by conventional methods. The following procedures used to fill in the gaps.

1. Filling by the average values of the original data.

2. Filling without selection, when a gap is filled with a constant value from an external source, such as a previous observation value from the same survey. The data obtained is generally considered to be a complete sample, i.e. the consequences of filling are ignored.

3. Filling with the selection, when substituted the values of variables from other objects in the sample. The substitution selected for each missing value according to the distribution estimate, as opposed to filling in the gaps with averages when the distribution average substituted. In most programs, the empirical distribution is defined by the values present, so different values from the data for similar objects are substituted when filling in with matching.

4. Filling by regression consists in filling in the missing values predicted by the regression of the missing variables for this object to the present ones calculated for complete objects. Assume that the variables *Ej* and *Ek* strongly correlated. The observation of a variable *Ej* is missing (or skipped) the *i*-th observation *eij* . It is natural to try to predict the *eij* value by *Ek* (or rather by *eik* observation) and then include this substitution (filled-in omission) in the analysis by *Ej* variable.

Weighting methods. Randomized conclusions from a sample of study gaps usually build on the weights of plan which, inversely proportional to the probabilities selected. Weighting is related to filling with average values according to the formula:

$$W_i = \frac{\sum p_i^{-1} x_i}{\sum p_i^{-1}}, \qquad (1)$$

where *xi* – value of the *i*-th sample variable *X*; *pi* – probability of extraction of the *i*-th sample variable in *X*. Weighting methods measure *Wi* weights (1) to account for the significance of the extracted variable. For example, let the plan weights constant in the sample subgroups. Then filling in the gaps in each subgroup with the mean of the subgroup and weighing the present values by their proportion in each subgroup leads to the same estimates of the mean from the sample, although the estimates of the sample variance are different, unless adjustments used to fill. Details of weighting methods described in l in [6].

Modeling-based methods. These based on the construction of the model of generation of gaps in the data. The conclusions obtained with the help of the likelihood

function, constructed under the condition of the validity of this model, with the estimation of the parameters by methods of the maximum likelihood type.

The advantage of these methods is that they are flexible, allow you to abandon the methods developed for special cases of omissions, and work with incomplete data of various kinds of samples on the common approach to maximizing the likelihood function.

It should noted that the choice of a particular method of data recovery with gaps in the data determined by the characteristics of both the data and the methods of obtaining them. In particular, for data processing with omissions in the form of binary tables, in which the elements of the tables take the values 1 or 0 ("+" or "-"), the method of homogeneous groups for the recovery of binary data proposed below, which is a modification of the known methods of filling [6,7] and the authors ' article [8].

Problem statement

Let we ask 10 possible trade experts for information about the evaluation of the proposed product in order to assess the consumer demand for it on 6 features (criteria) [1]. The survey data summarized in table 1. Here "+" – a positive answer, «-» – a negative answer, «?» – lost answer due to the fault of or the subject of the survey, or for technical reasons, the answer, which will be denoted by a question mark and interpreted as a «pass» of the data. It is necessary to restore the missing binary data for one of the above reasons to assess the consumer demand of survey subjects for the selected parameters. The solution of this problem proposed to carry out on the method of homogeneous groups.

Problem statement

Let we ask 10 possible trade experts for information about the evaluation of the proposed product in order to assess the consumer demand for it on 6 features (criteria) [1]. The survey data summarized in table 1. Here "+" – a positive answer, «-» – a negative answer, «?» – lost answer due to the fault of or the subject of the survey, or for technical reasons, the answer, which will be denoted by a question mark and interpreted as a «pass» of the data. It is necessary to restore the missing binary data for one of the above reasons to assess the consumer demand of survey subjects for the selected parameters. The solution of this problem proposed to carry out on the method of homogeneous groups.

Homogeneous groups method

The essence of this method is to find in the general population of binary data homogeneous groups of subjects of the survey and determine the affiliation of the subject of the survey, in the response of which for one of the above reasons there was omission in the data, to one of the selected homogeneous groups. Further, to recover the lost information,

you can use the traditional filling methods described above. We show a practical implementation of the homogeneous group method for the above problem statement.

As can be seen from Table 1, for the subjects of the survey i =2.7 the answer to the question about the evaluation of the trait i = 6 is lost, i.e. there are omissions.

Table 1. **The results of the survey**

	j	**Subjects of the survey**									
№п/п	**i**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Characteristics of the good	1	-	+	+	+	+	-	-	+	-	-
	2	+	+	+	+	+	-	-	+	+	+
	3	-	-	-	-	+	+	+	-	+	+
	4	-	+	+	+	-	-	+	+	+	-
	5	+	-	-	-	+	-	+	-	-	-
	6	-	?	+	+	+	-	?	-	-	-

For each of these subjects we will form groups with the same estimates for all the features (homogeneous groups), giving the values of the omission «+» and «-» – sequentially, and denote the number of subjects in these groups through r, and the number of identical features through S.

Analysis of the data in Table 1 shows that for the subject of the survey 2 (i = 2), if we take sign 6 (j = 6) as «+», then r = 3, S = 4. At the same time, a homogeneous group consists of the subjects of the survey 2, 3, 4. If we take sign 6 (j = 6) as «-», then r = 1, S = 6. Reasoning similarly for the subject of the survey 7 (i = 7), we obtain in the case of «+» r = 1, S = 4 , and in the case of «-» r = 1, S = 4. It should noted that a homogeneous group in both cases consists of one subject of the survey 7 (j = 7).

We introduce a generalized measure of the number of matched estimates in the cases of «+» and «-» in each homogeneous group

$$N = rS, \tag{2}$$

the maximum of which determines pass sign.

Hence, according to (2), for the subject of the survey 2 (i = 2) for the value «+» we get N = 12 , and for the value «-» we get N = 6 . For the subject of the survey 7 (i = 7) for the value «+» we get N = 4, and for the value «-» we get N = 6.

Then, according to one of the methods of filling, namely the replacement method [6,7], we can put in place a pass for i = 2 – «+», for i = 7 – «-», since it is for these values in the considered subjects with omissions that the generalized indicator of the number of matched estimates has the maximum value. As a result, table 1 will take the form of the final evaluation of the goods in the form of Table 2.

 ISSN 1560-8956

Further, summing up the number of positive and negative answers for each line, we can conclude what feature (or characteristic) of the goods needs to be improved to increase consumer demand [1], in particular, in this case, these are the features of the goods 3, 5, 6.

Table 2. **The final evaluation of the good**

	j	Subjects of the survey									
№п/п	i	1	2	3	4	5	6	7	8	9	10
Characteristics of the good	1	-	+	+	+	+	-	-	+	-	-
	2	+	+	+	+	+	-	-	+	+	+
	3	-	-	-	-	+	+	+	-	+	+
	4	-	+	+	+	-	-	+	+	+	-
	5	+	-	-	-	+	-	+	-	-	-
	6	-	+	+	+	+	-	-	-	-	-

Conclusion

This method works well for large amounts of binary data with a relatively small number of gaps, but, like all methods of filling, has a number of the above disadvantages and is used where the gaps are not critical. Otherwise, stricter methods are used, for example, modeling methods using maximum likelihood functions or Bayesian strategies [5,6].

REFERENCES

1. Aivazyan S. A., Enyukov I. S., Meshalkin L. D. Applied statistics. Fundamentals of modeling and primary data processing-M.: Finance and statistics. 1983.– 472p.

2. Zozulev A. V., Solntsev S. A. Marketing research: Theory, methodology, statistics: a tutorial.-Kyiv. Znannya, 2008. – 643p.

3. Doctorov B. Z. Post-Gallup poll technologies: To the 200th anniversary of public opinion polls in the United States // Sociological journal. 2005. №. 2. pp. 5-36.

4. Anderson T. Introduction to multivariate statistical analysis – M.: Fizmatgiz. 1963.– 499p.

5. Borovkov A. A. Mathematical statistics. M.: Science. 1984. –472p.

6. R. J. A. little, D. B. Rubin, Statistical analysis with missing data. – M.: Finance and statistics. 198.1– 336p.

7. Kohren U. Methods of sample study. M.: Statistics.-1976. – 440p.

8. Gubsky A. N., Stenin A. A., Korchinsky V. M. Method of binary data recovery with omissions //System technologies - Dnipro: NMetAU, IVK "System Technologies",№1(96) 2015. – pp. 206-211.