# STATISTICAL DECISIONS IN PRESENCE OF IMPRECISELY REPORTED ATTRIBUTE DATA

Olgierd Hryniewicz
*Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland*

Keywords: Statistical decisions, Attribute data, Imprecise data, Fuzzy approach, Possibility distribution.

Abstract: The paper presents a new methodology for making statistical decisions when data is reported in an imprecise way. Such situations happen very frequently when quality features are evaluated by humans. We have demonstrated that traditional models based either on the multinomial distribution or on predefined linguistic variables may be insufficient for making correct decisions. Our model, which uses the concept of the possibility distribution, allows to separate stochastic randomness from fuzzy imprecision, and provides a decision – maker with more information about the phenomenon of interest.

## 1 INTRODUCTION

When we make decisions on the basis of statistical analysis of data, we call such decisions – statistical decisions. An important class of statistical decisions is based on attribute data. In the simplest case, attribute data are presented in a form of a random sample consisting of elements having only two values: zero and one. All cases described by zeroes are usually called "failures", and the remaining statistical observations are called "successes".

Statistical decisions based on attribute data are well known in many fields of application. They were introduced more than eighty years ago in statistical quality control, and since that time they have been widely used in industry, business and administration. However, in information technology these methods are, as for know, not very popular. Take, for example, typical decision problems of Artificial Intelligence or Pattern Recognition. Quality of proposed algorithms is evaluated on widely accepted benchmarks without taking into account the randomness of their outputs which results from the randomness of input data. In this paper we present an attempt to deal with this problem in cases which seem to be typical in such applications like e.g. linguistic summarizations of text data or automatic classification of documents.

The theory of statistical decisions for the attribute data (i.e. 0 – 1) is well known for more than eighty years. It has been developed mainly for applications in statistical quality control or other industrial applications. In all such cases each element of the analysed sample is precisely evaluated as either "success" (1) or "failure" (0). However, in many areas of application such precise evaluations are hardly possible. Consider, for example, an automatic selection of text documents, where users evaluate the appropriateness of the selection. The proportion of documents which have been wrongly classified may serve as a measure of the effectiveness of this algorithm. In many cases however, it is difficult to present unequivocal evaluations. The users may prefer to have a possibility to give also answers like "May be Yes", "I am Undecided" or "May be Not", and not only either "Yes" or "No". To give another example from the area of information technology, let us consider the evaluation of a new algorithm for the compression of graphics. The perceived quality of this new method can be evaluated by a group of experts who are asked about the acceptability of compressed pictures.

The practical necessity to work with such imprecisely reported data prompted some authors to develop appropriate statistical tools that could be useful in decision making. The simplest approach is based on the multinomial model for imprecisely reported attribute data. We present this model in the second section of the paper. Another possibility, the application of fuzzy linguistic variables is analysed in the third section. In the fourth section we present a new approach based on the possibility theory introduced by Zadeh. We present a possibilistic

Hryniewicz O. (2009).
STATISTICAL DECISIONS IN PRESENCE OF IMPRECISELY REPORTED ATTRIBUTE DATA.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 309-312
DOI: 10.5220/0001863003090312

generalization of the multinomial model. In Monte Carlo simulation experiments which are not described in this short paper we have shown that this approach provides more information for decision makers in comparison to the aforementioned methods.

## 2 MULTINOMIAL MODEL FOR IMPRECISE ATTRIBUTE DATA

Suppose that a random variable, representing statistical data of interest, may have $k$ distinct values. These values can be represented by natural numbers, but can be also represented by either ordered or unordered labels. The probabilities of observing those values are denoted by $(p_1,\ldots,p_k)$, where $\sum_{i=1}^{k} p_i = 1$. If we observe a random sample of $n$ realization of this random variable, the probability distribution that describes the numbers of occurrences of all possible values $(X_1,\ldots,X_k)$ of this random variable is called the multinomial distribution, and is defined by the following function

$$P(X_1 = x_1,\ldots,X_k = x_k) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k} \sum_{i=1}^{k} x_i = 1. \quad (1)$$

This distribution may be used for the construction of decision-making procedures when observed values may be assigned to $k$ different categories. For example, in statistical quality control we may observe different types of failures. If for each considered type of failure we fix a critical number of nonconforming items in the sample, we can use (1) for the calculation of the probability of the acceptance of the sampled population for all possible values of the probabilities $(p_1,\ldots,p_k)$.

Let us now consider the situation when observed attribute data are imprecisely described by linguistic labels. Without loosing generality, we may assume that these data are described by the following set of labels: "Yes" (Y) "May be Yes" (MY), "I am Undecided" (U), "May be Not" (MN), and "Not" (N). Let denote by $(p_Y, p_{MY} \ldots, p_N)$ the vector of the corresponding probabilities of observations. Then, we can use (1) for the calculation of all interesting probabilities. We should note, however, that in the considered case the decision – making procedure should be different than in the aforementioned case of statistical quality control. We are usually interested in the unknown proportion of actual (A) successes. Let us assume that we may make only two decisions: "Accept" (if the actual proportion of "successes" is small) or "Reject" (if otherwise). The decision is based on the number of "successes" in the sample. If this number is not greater than a certain critical number $c$ our decision is to "Accept". Otherwise, the decision is to "Reject". The decision criterion $c$ is determined from the analysis of the probability of "Acceptance" calculated from the appropriate binomial distribution.

In the considered case of imprecisely reported observations actual "successes" may be hidden under four possible labels, i.e Y, MY, U, and even MN. Therefore, we may think about four possible critical numbers: $c_Y$, for observations, which occur with probability $p_{(Y)}=p_Y$, $c_{MY}$, which occur with probability $p_{(MY)}=p_Y+p_{MY}$, $c_U$, which occur with probability $p_{(U)}=p_Y+p_{MY}+p_U$, and $c_{MN}$, which occur with probability $p_{(MN)}=p_Y+p_{MY}+p_U+p_{MN}$. In order to set all these critical values we have to know all acceptable values for all these probabilities. However, in practice we know these values only for the actual probability of a "success". Therefore, it is natural to use only one critical value $c$ for all these possible outcomes of the test. If we do so, it is easy to show that the probabilities of "Acceptance" will be quite different, depending on the values of the probabilities $(p_Y, p_{MY} \ldots, p_N)$. However, usually we do not know these probabilities, so we don't know the actual characteristics of our decision procedure. Therefore, the multinomial model, if it has to be used for the modelling of imprecise attribute data, requires additional knowledge about the probabilities of different answers.

## 3 FUZZY LINGUISTIC VARIABLES AS MODELS OF IMPRECISE ATTRIBUTE DATA

Imprecise values of attribute data can bee looked upon as linguistic data described by fuzzy linguistic variables as it was proposed by Zadeh. For modelling quality data his approach was adopted in (Wang and Raz, 1990) who proposed to describe imprecise answers by predefined fuzzy subsets of the interval [0,1]. In their original paper they proposed to use fuzzy triangular number defined on overlapping subsets of [0,1]. For making decisions Raz and Wang proposed to use some real-valued representations of fuzzy numbers, such as: modal

value, midpoint of the 50% $\alpha$-cut, average, and centroid. It is easy to show that for the first three of the abovementioned representations it does not matter if we calculate representative values for individual observation and then sum them up or if we calculate a fuzzy sum, and then its representative value. In the fourth case this important property holds only either for triangular fuzzy numbers or for rectangular fuzzy numbers (i.e. for intervals).

In order to make decision about "Acceptance" (or "Rejection") we have to compare the representative value of the sum of observed linguistic variables with a certain critical value. Unfortunately, this critical value cannot be easily calculated for a simple reason that the representative values of the fuzzy sum of fuzzy observations may be quite different from the expected number of evaluated "successes". Especially when the fraction of imprecise observations is significant the observed representative values may be quite different than the expected numbers of "successes" in the sample.

Another problem with determination of a correct critical value for representative values of fuzzy observations is related to their strong dependence on the assumed representations of imprecise linguistic concepts. All these problems and difficulties make decision – making which is based on this fuzzy approach rather questionable.

It is also worth noticing that in all cases when calculation of representative values can be performed on individual fuzzy observations the whole procedure boils down to ordinary weighting of observations. This concept is also known as the calculation of "demerits", and has been successfully implemented in statistical process control (SPC). However, in SPC it is assumed that available information let us compute probabilistic characteristics of the considered statistic. Unfortunately, this is usually not the case for the problem considered in this paper. Recently, in (Gülbay and Kahraman, 2007) another fuzzy approach has been proposed for the analysis of linguistic quality data. However, this approach in the context of decision-making has exactly the same limitations as that of Wang and Raz.

# 4 POSSIBILITIC MODEL OF IMPRECISE ATTRIBUTE DATA

In the previous two sections we have demonstrated that in case of imprecisely reported attribute data the information provided in terms of simple linguistic labels may be not sufficient for correct decision – making if this correctness should depend upon the fraction of "successes" in a considered population. In (Hryniewicz, 2008) an extension of the considered model has been proposed by allowing additional information about imprecise observations. Our extension is based on a fact that each observation may be treated as a "success", but to a certain degree, and vice versa, as a "failure", but also to a certain degree. Thus, the result of each observation can be described by a fuzzy set

$$\mu_0 \,|\, 0 + \mu_1 \,|\, 1,\ 0 \le \mu_0, \mu_1 \le 1,\ max\{\mu_0, \mu_1\} = 1, \qquad (2)$$

defined on the set $\{0,1\}$. This fuzzy representation may be also interpreted as a possibility distribution over the set of two crisp outcomes of an observation: "success" (one) and "failure" (zero). When the result of an observation is described linguistically in such a way that it can be regarded as a "failure", the result of observation is expressed as a fuzzy set with the membership function $1 \,|\, 0 + \mu_1 \,|\, 1$. Full (i.e. undoubted) "failures", which in our setting are represented by labels "No", are now described by crisp sets. In this case the membership function is given by $1 \,|\, 0 + 0 \,|\, 1$. When $0 < \mu_1 < 1$ the corresponding label is "May be No", and $\mu_1$ in this case describes the degree to which this label is incompatible with an unequivocal label "No". On the other hand, if the result of an observation is described linguistically in such a way that it can be regarded as a "success", the result of observation is expressed as a fuzzy set with the membership function $\mu_0 \,|\, 0 + 1 \,|\, 1$. Full (i.e. undoubted) "successes", which in our setting are represented by a labels "Yes", are described by crisp sets with the membership function $0 \,|\, 0 + 1 \,|\, 1$. When $0 < \mu_0 < 1$ the corresponding label is "May be Yes", and $\mu_0$ in this case describes the degree to which this label is incompatible with an unequivocal label "Yes". When $\mu_0 = \mu_1 = 1$, we have the situation which we describe by a label "Undecided", as in this case there is the same possibility either of "successes" and "failures".

Assume now, that in the sample of $n$ items $n_1$ cases are characterized by fuzzy sets described by the membership function

$$\mu_{0,i} \,|\, 0 + 1 \,|\, 1,\ i = 1, \ldots, n_1, \qquad (3)$$

and in the remaining $n_2 = n - n_1$ cases by fuzzy sets described by the membership function

$$1 \,|\, 0 + \mu_{1,i} \,|\, 1,\ i = 1, \ldots, n_2. \qquad (4)$$

Without loss of generality we can assume that

$$0 \le \mu_{0,1} \le \ldots \le \mu_{0,n_1} \le 1, \quad (5)$$

and

$$1 \ge \mu_{1,1} \ge \ldots \ge \mu_{1,n_2} \ge 0. \quad (6)$$

Hence, the fuzzy total number of "successes" in this sample, calculated using Zadeh's extension principle, is given by:

$$\begin{aligned} \widetilde{x} = \quad & \mu_{0,1} \,|\, 0 + \mu_{0,2} \,|\, 1 + \ldots + 1 \,|\, n_1 + \\ & + \mu_{1,1} \,|\, (n_1 + 1) + \ldots + \mu_{1,n_2} \,|\, (n_1 + n_2). \end{aligned} \quad (7)$$

This number has to be compared with a critical number of "successes" $c$ in order to make a decision of "Acceptance" or of "Rejection".

Comparison of fuzzy numbers cannot be done unequivocally, as they are not completely ordered. One of the widely accepted methods of comparison is based on the concepts of possibility and necessity of dominance introduced in (Dubois and Prade, 1983). Let $\mu(x)$ be the membership function of the fuzzy set $\widetilde{X}$, and $\nu(y)$ be the membership function of the fuzzy set $\widetilde{Y}$. When the evidence that $\widetilde{X}$ is strictly greater than $\widetilde{Y}$ is rather strong we can express this feature using the *Necessity of Strict Dominance (NSD)* index, defined as follows:

$$NSD\left(\widetilde{X} \succ \widetilde{Y}\right) = 1 - \sup_{x,y:x \le y} \left[\min\left(\mu(x), \nu(y)\right)\right]. \quad (8)$$

When this evidence is weak, we can use the *Possibility of Strict Dominance (PSD)* index which is related to the *NSD* using the following formula.

$$NSD\left(\widetilde{X} \succ \widetilde{Y}\right) = 1 - PSD\left(\widetilde{Y} \succ \widetilde{X}\right). \quad (9)$$

The interpretation of these indices reflects common understanding of the words "possible" and "necessary". Relations which are only partially possible ($PSD < 1$) are not necessary ($NSD = 0$). On the other hand, relations which are even partially necessary ($NSD > 0$) are always fully possible ($PSD = 1$).

It is easily seen that when the number of "successes" with the maximal value of the membership function (equal to one) is situated to the left of $c$ we can say about a certain necessity that the relation $x<c$ has been fulfilled. This necessity is equal to one only in case when the whole support of $\widetilde{x}$ (i.e. values of $x$ with positive membership) is located to the left of $c$. When the number of "successes" with the maximal value of the membership function (equal to one) is situated to the right of $c$ we can only say about a certain possibility that the relation $x<c$ has been fulfilled. This possibility is equal to zero only in case when the whole support of $\widetilde{x}$ is located to the right of $c$.

This interpretation of possibility and necessity indices let us formulate simple rules for decision – making. We have to fix the critical value $c$ and the required value of the necessity/ possibility of $\widetilde{x} \prec c$. Thus, we are able to define a non-fuzzy decision rule.

# 5 CONCLUSIONS

In the paper we have proposed a new approach to the analysis and decision – making when information is presented in a form of imprecisely reported attribute data. We have demonstrated on examples that traditional and popular approaches provide only restricted information which might be insufficient for correct decision making. The new approach is definitely more flexible. Moreover, it can be straightforwardly extended to the case when the definitions of "success" and "failure" may be imprecise. This imprecision may lead to imprecise (fuzzy) decision criteria.

# REFERENCES

Dubois, D., Prade, H., 1983. Ranking Fuzzy Numbers in the Setting of Possibility Theory. *Information Science,* vol.30, 183-224.

Gülbay, M., Kahraman, C., 2007. An alternative approach to fuzzy control charts: Direct fuzzy approach. *Information Sciences,* vol.177, 1463-1480.

Hryniewicz, O., 2008. Statistics with fuzzy data in statistical quality control. *Soft Computing*, vol.12, 229-234.

Wang, J.H., Raz, T., 1990. On the Construction of Control Charts Using Linguistic Variables. *International Journal of Production Research,* vol.28, 477-487.