The Mathematics behind *WizWhy*

This appendix briefly presents the formulas used to calculate some of the figures presented by *WizWhy*. The appendix is aimed at those who are interested in mathematics. If you are not, simply skip it. You don't have to be a mathematician in order to use *WizWhy* effectively.

The error probability of an if-then rule

We will use the following notations:

m is the number of cases in the if-then rule;
n is the number of records satisfying the rule's condition;
N is the total number of records in the investigated file;
M is the number of records where the dependent variable R is r;
α is an error probability of this rule.

α is calculated as follows:

$$\alpha = \sum_{k=m}^{n} P_{N,M}(n,k),$$

where

$$P_{N,M}(n,k) = \frac{\binom{k}{M}\binom{n-k}{N-M}}{\binom{n}{N}}$$

The error probability α for the if-then-not rule is calculated as follows:

$$\alpha = \sum_{k=0}^{m} P_{N,M}(n,k)$$

Unexpected rules

Definition: A rule containing q-conditions ($q > 1$) in the "if" part, *R is r* in the "then" part, and having probability P that *R is r* is called *unexpected* if at least one of the following two requirements is fulfilled:

(1) There is a rule containing *R is not r* in the "then" part and q_1 -conditions ($q_1 < q$) in the "if" part such that the set of 1-conditions entered in the q_1-condition is a subset of the set of 1-conditions entered in the q-condition.

(2) There is no rule containing *R is r* in the "then" part and q_1-conditions ($q_1 < q$) in the "if" part (where the set of 1-conditions entered in the q_1-conditions is a subset of the set of 1-conditions entered in the q-conditions) for which the inequality $p << P$ is not fulfilled, where p denotes a probability of such a rule. (If $r = 1$, p is a probability that R is 1; if $r = 0$, p is a probability that R is 0.)

The second requirement means that the probability that R is r for any discovered rule with the above-defined q_1-conditions in the "if" part and R is r in the "then" part must be much less than P.

The unexpected rule parameters

Unlike other if-then rules, each Unexpected Rule revealed by *WizWhy* contains the following additional parameters:

1. Expected rule's probability
2. Level of unlikelihood

We will use the following notations:

N is the total number of records in the investigated file;
M is the total number of records where the Field to Predict R is r ;
p_a is the a priori probability that $R = r$, that is, $p_a = \frac{M}{N}$;
m_i is the number of records satisfying the ith 1-condition entered in the q-condition of the unexpected rule, $i = 1,\ldots,q$;

S_i is the number of records satisfying both the ith 1-condition and the condition R is r , $i = 1,\ldots,q;$;

p_i is the probability that R is r under the ith 1-condition, $p_i = \frac{s_i}{m_i}$

K is the number of records satisfying the q-condition of the unexpected rule;

S is the number of records satisfying both the q-condition of the unexpected rule and the condition R is r ;

$P_{\exp}$ is the expected probability for the considered unexpected rule;

U is the level of unlikelihood.

The *expected probability* $P_{\exp}$ is calculated as follows:

$$P_{\exp} = \frac{P_{dep} - P_{ind}}{k_{dep} - k_{ind}} \cdot (K - k_{ind}) + P_{ind} \ , \qquad \text{if} \quad k_{ind} \le K \le k_{dep} \ ;$$

$$P_{\exp} = P_{ind} \ , \qquad \text{if} \quad K < k_{ind} \ ,$$

where:

$$P_{dep} = \frac{\sum_{i=1}^{q} p_i}{q} \ ;$$

$$k_{dep} = \min_{i=1,\ldots,q} m_i \ ;$$

$$P_{ind} = \frac{1}{1 + \left(\frac{p_a}{1 - p_a} \right)^{q-1} \prod_{i=1}^{q} \frac{1 - p_i}{p_i}} \ ;$$

$$k_{ind} = \frac{\prod_{i=1}^{q} m_i}{N^{q-1}}$$

The *level of unlikelihood* U is calculated as follows:

$$U = \max_{i=1,\ldots,q} (1 - U_i) \ ,$$

where

$$U_i = \frac{\binom{S}{s_i} \cdot \binom{K-S}{m_i - s_i}}{\binom{K}{m_i}}$$

Prediction

When *WizWhy* issues a prediction on the basis of the if-then rules (and not the if-and-only-if rules), and the dependent variable was analyzed as Boolean, *WizWhy* lists the following parameters.

1. Weighted average rule probability
2. Prediction's significance level: error probability
3. Primary prediction's probability
4. Conclusive prediction's probability

All the rules applicable to the record under analysis are considered for calculating these parameters. Assume that n such rules have been found: the first k rules are of the "if-then" type, and the remaining $n-k$ rules are of the "if-then-not" type. For rule i let us denote the rule's probability, the error probability, and the number of cases in the rule by p_i, α_i, and m_i respectively, $i = 1,\ldots,n$.

The *weighted average rule probability* is calculated as follows: The significance level of rule i *is* taken as a weight for this rule's probability. The significance level of rule i is equal to $1-\alpha_i$. However, if the significance level is equal to 0.5, then it may be assumed that a measure of our confidence in the rule's validity is near to zero. Therefore, to evaluate the weight of the rule's probability, it is worthwhile to map the set of significance level's values changed within the segment [0.5 , 1] into segment [0 , 1]. Assume that this mapping is linear, and for $\alpha_i = 0$, the weight is equal to 1; for $\alpha_i = 0.5$ the weight is equal to 0. Then the weight

for the rule's probability with a priori probability α_i is equal to $1-2\alpha_i$. Thus, the weighted average rule probability P is:

$$P = \frac{\sum_{i=1}^{k}(1-2\alpha_i)\cdot p_i + \sum_{i=k+1}^{n}(1-2\alpha_i)\cdot(1-p_i)}{\sum_{i=1}^{n}(1-2\alpha_i)}$$

The *primary prediction's probability* is the average probability that the value of the dependent variable R is the predicted value r. The primary probability is presented to enable the user to evaluate the difference between the prediction's probability and the average probability. Let M be the frequency of value r of field R within the investigated file. Let N be the number of the investigated file's records. Then the primary probability P_a is equal to $\frac{M}{N}$.

The *prediction's significance level* means the measure of confidence that the prediction's probability P is true. The significance level is equal to $1-\alpha$, where α is the prediction's *error probability*. Error probability α is calculated as follows.

Assume that $P \geq P_\alpha$. The probability α_{pos} that at least one of k if-then rules is not erroneous is first calculated.

$$\alpha_{pos} = 1 - \prod_{i=1}^{k}\alpha_i$$

The probability α_{neg} that at least one of $n-k$ if-then-not rules is not erroneous is calculated as follows:

$$\alpha_{neg} = 1 - \prod_{i=k+1}^{n}\alpha_i$$

We assume that all the discovered rules are independent; that is, the complete set of events (cases) defined by all the rules is a union of non-intersecting sets, each of which is the set of events (cases) of a certain rule. Then, the error probability for the prediction is calculated as follows:

$$\alpha = 1 - \frac{\alpha_{pos} \cdot \sum_{i=1}^{k} n_i \ + (1 - \alpha_{neg}) \cdot \sum_{i=k+1}^{n} n_i}{\sum_{i=1}^{n} n_i} \ ,$$

where $n_i = \frac{m_i}{p_i}$.

If $P \geq P_a$, then α_{pos} is calculated for if-then-not rules, and α_{neg} is calculated for if-then rules.

The *conclusive prediction's probability* is the mathematical expectation *MP* of a probability *P* for the prediction. It is calculated as follows:

$$MP = (1 - \alpha) \cdot P + \alpha \cdot P_{\alpha} \ ,$$

where:

P is the weighted average rule probability
α is the prediction's error probability
P_a is the primary prediction's probability

The deduction of the formula can be explained as follows.

Consider the aggregation of all the relevant rules as the certain "generalized" rule. Let *A* be the following event: according to the given values of condition fields, this "generalized" rule has been found. The only parameter describing this event is the probability *P* that *R* = *r* (weighted average rule probability). We will consider this parameter as a random variable since we have found not only the value *P* of the parameter, but also the probability (equal to the "generalized" rule's significance level $1-\alpha$) of this value.

Consider the complete space of events as follows:

A is the event that the "generalized" rule has been found;

Ā is the event that the "generalized" rule has not been found.

The value of the above parameter (the probability that *R is r*) for the event Ā can be determined as follows: If the "generalized" rule was not found,

then this parameter's value is equal to the primary probability P_a. Since the event $\bar{A}$ is the complement of event A with respect to the whole space of events, the probability of value P_a is equal to α. The formula follows from the definition of a mathematical expectation of the considered random variable.