*Int. J. Appl. Math. Comput. Sci., 2012, Vol. 22, No. 3, 629–645*
DOI: 10.2478/v10006-012-0048-z

# OPTIMAL ESTIMATOR OF HYPOTHESIS PROBABILITY FOR DATA MINING PROBLEMS WITH SMALL SAMPLES

ANDRZEJ PIEGAT [*], MAREK LANDOWSKI [**]

[*] Faculty of Computer Science
West Pomeranian University of Technology, Żołnierska 49, 71-210 Szczecin, Poland
e-mail: apiegat@wi.zut.edu.pl

[**] Institute of Quantitative Methods
Maritime University of Szczecin, Wały Chrobrego 1–2, 70-500 Szczecin, Poland
e-mail: m.landowski@am.szczecin.pl

The paper presents a new (to the best of the authors' knowledge) estimator of probability called the "$Ep_{h\sqrt{2}}$ completeness estimator" along with a theoretical derivation of its optimality. The estimator is especially suitable for a small number of sample items, which is the feature of many real problems characterized by data insufficiency. The control parameter of the estimator is not assumed in an *a priori*, subjective way, but was determined on the basis of an optimization criterion (the least absolute errors).The estimator was compared with the universally used frequency estimator of probability and with Cestnik's $m$-estimator with respect to accuracy. The comparison was realized both theoretically and experimentally. The results show the superiority of the $Ep_{h\sqrt{2}}$ completeness estimator over the frequency estimator for the probability interval $p_h \in (0.1, 0.9)$. The frequency estimator is better for $p_h \in [0, 0.1]$ and $p_h \in [0.9, 1]$.

**Keywords:** single-case problem, probability, probability estimation, frequency interpretation of probability, completeness interpretation of probability, uncertainty theory.

## 1. Introduction

Probability is a very important form of uncertainty description, though not the only. Other alternative forms are fuzzy systems of Zadeh (Klirr and Yuan, 1996; Zadeh, 1965), Dempster–Shafer belief/plausibility theory (Shafer, 1976), possibility theory of Dubois and Prade (1988), and info-gap theory of Ben-Haim (2006). However, probability theory seems to be most important in this group of sciences. Thousands of students all over the world acquaint themselves with it.

It is used in derivation of many mathematical formulas applied in physics, measurement theory, statistics, various identification methods, and in artificial intelligence, e.g., in a probabilistic version of rough and fuzzy set theory (Polkowski, 2002; Ziarko, 1999) for probability evaluation of decision rules, in classification, clusterization, data mining, machine learning (Witten and Frank, 2005), etc. Many of the methods are based on the assumption of a large number of sample items. However, in real problems, this assumption is frequently not satisfied. Even in the case when we possess a seemingly large number of sample items, when the input space was partitioned into influence subspaces of particular rules, the number of sample items being in a single rule subspace becomes frequently very small. Figure 1 presents an example of input space partition typical for rough set theory.

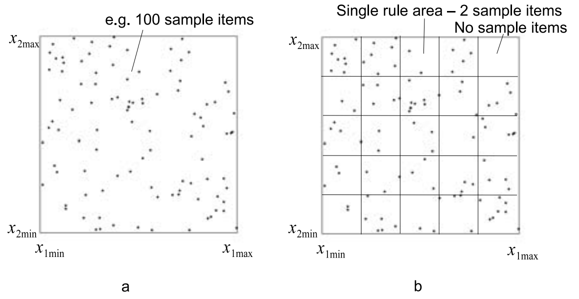


Fig. 1. Example illustrating a small number of sample items in the problem of probability estimation of rules detected with rough set theory or fuzzy set theory: sample items in the whole input space (a), sample items in the subspaces of single rules (regular input-space partition) (b).

A similar problem concerns also decision rules detected with the method of decision trees (Rokach and Maimon, 2008). In this case, influence subspaces of particular rules do not create the regular input-space partition as in the case of rough sets (Fig. 2).

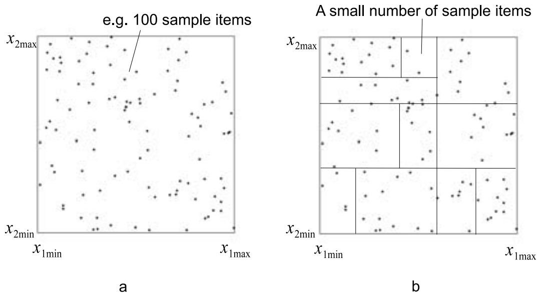


Fig. 2. Illustration of a small number of sample items occurring in influence subspaces of particular rules detected with the method of decision trees: sample items in the whole input space (a), sample items in subspaces of single rules (non-regular input space partition) (b).

The problem of a small number of sample items in influence subspaces of particular rules occurs frequently not only in rough set models, fuzzy set models or decision trees models. It also occurs in classification, clusterization, machine learning, data mining and in classic modeling problems.

Books mostly describe two main interpretations of probability: the classic and the frequency ones. The classic interpretation can shortly be presented as follows (Hajek, 2010). If a random experiment can result in $N$ mutually exclusive and equally likely outcomes and if $N_A$ of these outcomes result in the occurrence of the event $A$, the probability of $A$ is defined by (1),

$$P(A) = \frac{N_A}{N}. \tag{1}$$

The main creator of the classical definition was Laplace (1814). It can be applied only in problems where there is only a 'finite' number of possible outcomes, e.g., in the case of a dice, where six outcomes are possible. In such problems the universe $U_o$ of possible outcomes is fully known and contains a finite number of results. But there are experiments where the number of possible outcomes can be infinite. Then the universe $U$ of possible outcomes is open and can never be fully learned. Such an experiment is, e.g., tossing a coin until it lands head up. The 'frequentists' made an attempt to correct the classical definition. Their main representative was von Mises (1957). According to them, the "probability of an event is its relative frequency of occurrence after repeating a process a large number of times under similar conditions. If we denote by $n_A$ the number of occurrences of an event $A$ in $n$ trials, then, if

$$\lim_{n\to\infty} \frac{n_A}{n} = p, \tag{2}$$

we say that $P(A) = p$" (Hajek, 2010).

This interpretation is called the long-run sequence interpretation. Because in practice a very large (infinite) number of experiments cannot be realized, we have to use finite frequentism or a finite sequence interpretation, according to which the probability is calculated on the basis of data we have at our disposal. The definition of probability according to the finite-sequence interpretation is as follows: "the probability of an attribute $A$ in a finite reference class $B$ is the relative frequency of actual occurrence of $A$ within $B$" (Hajek, 2010). According to this interpretation, the probability estimate is calculated with the formula (3),

$$P(A) = \frac{n_A}{n}, \tag{3}$$

where $n$ is a finite number.

The frequentist interpretation of probability has many weak points. Scientists proved that it results in many illogicalities, paradoxes and unanswered questions (Burdzy, 2009; 2011a; Hajek, 2010; Piegat, 2011a; 2011b). The weak points and faults of frequentist interpretations were the reason for many scientists to search for new interpretations of probability. The most known alternative interpretations are subjective probability theory proposed by De Finetti (1975), the logical probability theory of Carnap (1952), and the propensity theory of probability of Popper (1957). Also Khrennikov (1999) and Rocchi (2003) proposed new interpretations. Readers can acquaint themselves with these interpretations in the works of Burdzy (2009; 2011a; 2011b), Hajek (2010) and Piegat (2011a; 2011b).

However, the new interpretations are also being discussed, and various questions and objections have been formulated against them. Therefore, some scientists are of the opinion that probability theory should be repulsed. One of them is Burdzy, whose book *The Search for Certainty. On the Clash of Science and Philosophy of Probability* (Burdzy, 2009) has aroused a vivid discussion among scientists (Burdzy, 2011a; 2011b).

Accepting many critical opinions concerning the universally used frequency interpretation of probability, Piegat proposed a completeness interpretation of probability (Piegat, 2011a; 2011b). Very shortly, according to this interpretation, to determine the probability of a hypothesis $h$ concerning an event, first the *complete evidential set* or, shortly, *Evidential Completeness (EC)* should be defined. It is a set of evidence pieces which would fully prove the truth of the hypothesis.

Because in practice we possess only a partial and incomplete evidence set, we can determine on its basis only the minimal, lower limit of probability $p_{h\,\min}$ of the hypothesis $h$ and the minimal probability $p_{\bar{h}\,\min}$ of the anti-hypothesis $\bar{h} = NOT\, h$. It enables calculation of the upper probability limits $p_{h\,\max}$ and $p_{\bar{h}\,\max}$ of the hypothesis and the anti-hypothesis, according to

$$\begin{aligned} p_{h\,\max} &= 1 - p_{\bar{h}\,\min}, \\ p_{\bar{h}\,\max} &= 1 - p_{h\,\min}. \end{aligned} \tag{4}$$

In most problems with probability forecasting in the open universe $U$ of events going outside the universe $U_o$ of the possessed evidence knowledge, we are not able to precisely determine the probability $p_h$ of the hypothesis concerning the event considered. Only an interval of its possible values (5) can be determined,

$$p_h \in [p_{h\,\min}, p_{h\,\max}] \tag{5}$$

Also the estimate $Ep_h$ of the probability $p_h$ can be determined, that is, its approximate value. However, the number of different estimators can be considerable. It depends on the estimation criterion and the way of estimation. Piegat (2011a; 2011b) proposed the first, simple completeness estimator $p_{hR}$. It represents the uncertainty interval (5) of the probability $p_h$ and it minimizes the maximal, absolute error of the estimate in relation to all possible values of $p_h$ (6),

$$p_{hR} = 0.5(p_{h\,\min} + p_{h\,\max}). \tag{6}$$

In the case of a binary event (e.g., a coin toss, occurrence of a certain event or not) in which $n_h$ means the number of confirmations of the hypothesis $h$ and $n_{\bar{h}}$ the number of confirmations of the anti-hypothesis $\bar{h}$, the estimator $p_{hR}$ takes the form of (7),

$$p_{hR} = \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2n_{SEC}}. \tag{7}$$

where $n_{SEC}$ means the number of results (evidence pieces) necessary for a satisfactorily precise (e.g., with accuracy of 99%) proof of the hypothesis truth. The details can be found in the works of Piegat (2011a; 2011b). Estimates $Ep_h$ calculated with the estimator (7) converge with an increasing number $n$ of evidence pieces to the precise value of the probability $p_h$. However, the convergence speed is small and can be increased. After many analyses, a new estimator of probability was found that will be demonstrated in the next section.

## 2. New estimator $Ep_{ha}$ of the probability $p_h$ of the hypothesis $h$

The universally used frequency estimator has many significant drawbacks. It has great errors for a small number $n$ of sample items with which we often have to deal in practical problems (data insufficiency). The frequency estimator also gives hardly acceptable results in the case of a single sample item. It suggests probability values $0$ or $1$, which means certainty. This phenomenon is called the “single case problem” (Burdzy, 2009; Hajek, 2010). When we have at our disposal only a homogeneous data set, e.g., $\{H, H, H, H, H\}$, where $H$ means, for instance the head of a coin, the frequency estimator also suggests the probability $p_h = 1$, which means certainty. This problem is better described by Piegat (2011b). The next fault of the frequency estimator is its considerable oscillating both at small and at large numbers of sample items where oscillations should not occur and the probability should stabilize (Burdzy, 2009; Larose, 2010).

In this paper a new estimator of the probability $p_h$ will be proposed that is generally denoted $Ep_{ha}$ (8),

$$Ep_{ha} = \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2(n+a)}, \tag{8}$$

where $n_h$ is the confirmation number of the hypothesis $h$, $n_{\bar{h}}$ is the confirmation number of the anti-hypothesis $\bar{h} = NOT\, h$ and $n = n_h + n_{\bar{h}}$ means the entire number of confirmations. A few examples of binary hypotheses and anti-hypotheses include the following:

- $h$: obesity increases the danger of diabetes, $\bar{h}$: obesity does not increase the danger of diabetes;
- $h$: in a coin the head dominates, $\bar{h}$: in a coin the tail dominates;
- $h$: speedy driving increases crash danger, $\bar{h}$: speedy driving does not increase crash danger.

Generally, in a problem more than two hypotheses relating to the problem outcome can be formulated. Then we speak about $n$-ary hypotheses, e.g., about trinary-hypotheses.

### Some general features of the $Ep_{ha}$ completeness estimator

• *Probability estimates calculated both by the completeness estimator $Ep_{ha}$ and by the frequency estimator $fr_h = n_h/n$ for a large number of sample items.* These estimates converge to the same value. It is the value of the true probability $p_h$ defined by (2). Below, the proof of this statement is given. Remark that $n = n_h + n_{\bar{h}}$, $a$ being a finite number.

We have

$$\begin{aligned} \lim_{n\to\infty} Ep_{ha} &= \lim_{n\to\infty} \left( \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2(n+a)} \right) \\ &= \lim_{n\to\infty} \frac{n + a + n_h - n_{\bar{h}}}{2(n+a)} \\ &= \lim_{n\to\infty} \frac{n_h + n_{\bar{h}} + a + n_h - n_{\bar{h}}}{2(n+a)} \end{aligned}$$

$$\begin{aligned}
&= \lim_{n\to\infty} \frac{2n_h + a}{2(n+a)} \\
&= \lim_{n\to\infty} \frac{2n_h}{2(n+a)} + \lim_{n\to\infty} \frac{a}{2(n+a)} \\
&= \lim_{n\to\infty} \frac{n_h}{n+a} = \lim_{n\to\infty} \frac{n_h/n}{1+a/n} \\
&= \lim_{n\to\infty} \frac{n_h}{n} = p_h.
\end{aligned} \tag{9}$$

Thus, the completeness estimator $Ep_{ha}$ for $n \to \infty$ identifies the precise value of the probability $p_h$, similarly to the universally used frequency estimator $fr_h$.

- The probability estimate $Ep_{ha}(1)$ calculated by the completeness estimator $Ep_{ha}$ from one single sample item.

If a single sample item is a confirmation of the hypothesis $h$ (the notation for such a sample item is $1_h$), then the estimate $Ep_{ha}(1_h)$ is determined by

$$Ep_{ha}(1_h) = \frac{1}{2}\left(\frac{2+a}{1+a}\right). \tag{10}$$

For $a > 0$, the estimate $Ep_{ha}(1_h)$ satisfies the condition $0.5 < Ep_{ha}(1_h) \leq 1$. If, e.g., $a = 1$, then $Ep_{ha}(1_h) = 0.75$. Instead, the probability estimate $p_h$ determined by the frequency estimator $fr_h(1_h)$ from one sample item confirming that the hypothesis $h$ equals 1,

$$fr_h(1_h) = \frac{n_h}{n} = \frac{1}{1} = 1. \tag{11}$$

Thus, it can be said that the frequency estimator realizes from one sample item drastic or extreme reasoning concerning the hypothesis $h$. If the single sample item is a negation of the hypothesis $h$ (denotation of such a sample item is $1_{\bar{h}}$), then the estimate $Ep_{ha}(1_{\bar{h}})$ is given by

$$Ep_{ha}(1_{\bar{h}}) = \frac{1}{2}\left(\frac{a}{1+a}\right). \tag{12}$$

If $a > 0$, then $Ep_{ha}(1_{\bar{h}})$ satisfies the condition $0 \leq Ep_{ha}(1_{\bar{h}}) < 0.5$. If, e.g., $a = 1$, then its value is given by

$$Ep_{ha}(1_{\bar{h}}) = 0.25. \tag{13}$$

In the case of the frequency estimator $fr_h(1_{\bar{h}})$, the probability estimate from one sample item $1_{\bar{h}}$ negating the hypothesis $h$ equals 0,

$$fr_h(1_{\bar{h}}) = \frac{n_h}{n} = \frac{0}{1} = 0. \tag{14}$$

Thus, also in this case the frequency estimator realizes drastic and extreme reasoning from one sample item. From a single negation of the hypothesis $h$, it concludes its zero-probability. Instead, reasoning about the probability $p_h$ realized by the completeness estimator $Ep_{ha}$ for $a = 1$ can be called a "moderate" one. In the case of this estimator, the reasoning caution (radicalism) can be controlled with coefficient $a$. For $a = 0$ (no caution), the estimator $Ep_{ha}$ becomes the frequency estimator $fr_h$ and realizes drastically radical and risky reasoning from one sample item. Increasing $a$ increases the reasoning caution from one sample item. As $a \to \infty$, the reasoning becomes maximally cautious (no risk, no radicalism).

- The probability estimate $Ep_{ha}(0)$ calculated by the completeness estimator $Ep_{ha}$ at the lack of sample items ($n = 0$).

The estimate value for this case is given by (15)

$$\begin{aligned}
Ep_{ha}(0) &= \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2(n+a)} \\
&= \frac{1}{2} + \frac{0-0}{2(0+a)} = \frac{1}{2}.
\end{aligned} \tag{15}$$

This estimate (it is a hypothesis referring to the real value, not a statement) is reasonable and acceptable because it minimizes to 0.5 the maximal possible absolute error of the estimate in relation to the real value of the probability $p_h$ in a situation when nothing is known about the probability. Any other estimate value $Ep_{ha}(0) \neq 0.5$ would increase the maximal possible estimate error over 0.5. Instead, the frequency estimator $fr_h$ at the lack of sample items ($n = 0$) is not able to infer any conclusion referring to the hypothesis probability $p_h$,

$$fr_h(0) = \frac{n_h}{n} = \frac{0}{0}, \tag{16}$$

because the result of division $0/0$ is undetermined.

## 3. Determining the optimal value of the coefficient $a$ of the reasoning caution from one sample item $1_h$ confirming the hypothesis $h$ for the completeness estimator $Ep_{ha}$

The coefficient $a$ in the formula (8) of the estimator $Ep_{ha}$ can be called a reasoning caution from one sample item because with its increase the caution increases while reasoning risk and radicalism decrease. Before deriving the formula for the optimal value of this coefficient, consider probability estimation from one, single sample item $1_h$. Let us assume that we have to deal with a coin for which the true head probability equals $p_h$ and the true tail probability equals $p_{\bar{h}} = 1 - p_h$. Let us assume that one toss gave the head, thus $n_h = 1$ and $n_{\bar{h}} = 0$. What will the frequency estimator conclude from this experiment result? It will conclude as below,

$$fr_h(1_h) = \frac{n_h}{n} = \frac{1}{1} = 1.$$

In most cases such a conclusion is incorrect because (in most cases, apart from the case $p_h = 1$) the true probability $p_h$ is not equal 1 but it has a fractional value, i.e., $0 < p_h \leq 1$.

Let us denote by $\Delta(1_h)$ the absolute error of such a conclusion,

$$\Delta(1_h) = |p_h - fr_h(1_h)| = |p_h - 1|. \tag{17}$$

Now, let us analyze the situation where a single toss did not give heads but tails ($n_h = 0$, $n_{\bar{h}} = 1$). Such a sample item will be denoted by $1_{\bar{h}}$ because it is a sample item confirming negation $\bar{h}$ of the hypothesis $h$ about head domination in the coin. Which conclusion concerning the hypothesis $h$ does the frequency estimator infer from such a sample item? We get

$$fr_h(1_{\bar{h}}) = \frac{n_h}{n} = \frac{0}{1} = 0.$$

Let us denote by $\Delta(1_{\bar{h}})$ the absolute error of such a conclusion,

$$\Delta(1_{\bar{h}}) = |p_h - fr_h(1_{\bar{h}})| = |p_h - 0| = p_h. \tag{18}$$

Now, let us assume that we have at disposal a large number $N \to \infty$ of sample items from experiments of coin tossing, and $N_h$ of these sample items are heads, which confirm the hypothesis $h$, and $N_{\bar{h}}$ sample items are tails, which confirm the anti-hypothesis $\bar{h}$ ($N = N_h + N_{\bar{h}}$). Let us denote by $\Delta(N_h)$ the sum of absolute errors of all individual conclusions from all $N_h$ single sample items,

$$\Delta(N_h) = N \cdot \Delta(1_h) = N \cdot |p_h - 1|. \tag{19}$$

Because the number $N$ of all sample items approaches infinity, then, according to the definition (2) of probability, $N_h = N \cdot p_h$ and $N_{\bar{h}} = N \cdot (1 - p_h)$. Thus, the error sum of all sample items confirming the hypothesis can be transformed from (19) into (20),

$$\Delta(N_h) = N_h \cdot |p_h - 1| = N \cdot p_h|p_h - 1|. \tag{20}$$

On the basis of a similar reasoning, we get the formula

$$\Delta(N_{\bar{h}}) = N_{\bar{h}} \cdot \Delta(1_{\bar{h}}) = N_{\bar{h}} \cdot p_h = N(1 - p_h)p_h \tag{21}$$

determining the error sum $\Delta(N_{\bar{h}})$ of all individual conclusions from sample items confirming the anti-hypothesis $\bar{h}$.

The error sum of conclusions from all $N$ sample items, both from $N_h$ sample items confirming the hypothesis $h$ and from $N_{\bar{h}}$ sample items confirming the anti-hypothesis $\bar{h}$, is determined by

$$\begin{aligned} \Delta(N) &= \Delta(N_h) + \Delta(N_{\bar{h}}) \\ &= Np_h|p_h - 1| + N(1 - p_h)p_h \\ &= 2N(1 - p_h)p_h. \end{aligned} \tag{22}$$

If $N$ approaches infinity, then so does the error sum $\Delta(N)$. It hampers theoretical analyses concerning the sum. However, we can calculate from (22) the mean, average conclusion error $\Delta^{aver}(1)$ of a single sample item, independently of whether it was the sample item $1_h$ confirming the hypothesis $h$ or the sample item $1_{\bar{h}}$ negating the hypothesis $h$,

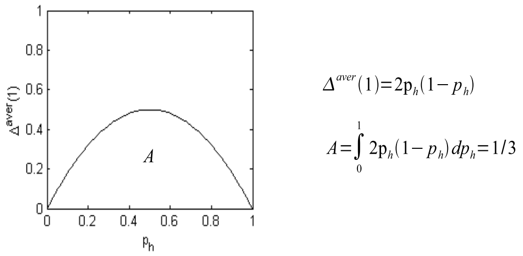


Fig. 3. Dependence of the average error $\Delta^{aver}(1)$ of the one-sample item conclusion concerning the real value $p_h$ of hypothesis probability for the frequency estimator $fr_h$.

$$\Delta^{aver}(1) = \Delta(N)/N = 2(1 - p_h)p_h. \tag{23}$$

The functional surface of the average error $\Delta^{aver}(1)$ of the reasoning about probability from a single sample item is demonstrated in Fig. 3.

Because in real problems we do not know the precise value of the probability $p_h$, inferring from one sample item we are convicted to make estimation errors of probability. A global measure of these errors is the area $A$ under the curve of $\Delta^{aver}(1) = f(p_h)$ that in case of the frequency estimator $fr_h$ in Fig. 3 equals $1/3$. The following question can be asked: "Does the application of other probability estimators, different from $fr_h = n_h/n$, allow decreasing the area $A$ of the average error $\Delta^{aver}(1)$ and thus decreasing errors of probability estimation?".

To answer this question, let us consider the idea described below.

Instead of the frequency estimator that from one sample item $1_h$ confirming the hypothesis calculates the estimate $fr_h = 1$ and from one sample item $1_{\bar{h}}$ negating the hypothesis calculates the estimate (conclusion) $fr_h(1_{\bar{h}}) = 0$, let us apply another estimator $Ep_h$ of a more cautious character, which from one single sample item $1_h$ will conclude (calculate) an estimate $Ep_h(1_h) \leq 1$ and from one sample item negating the hypothesis will conclude an estimate $Ep_h(1_{\bar{h}}) \geq 0$ satisfying the condition (24),

$$Ep_h(1_h) + Ep_h(1_{\bar{h}}) = 1. \tag{24}$$

It should be mentioned that (24) is satisfied also by the frequency estimator $fr_h$. Because the true but unknown hypothesis probability equals $p_h$, the error $\Delta^{aver}(1_h))$ of concluding from one sample item by the estimator $Ep_h$ is given by

$$\Delta(1_h) = |p_h - Ep_h(1_h)|. \tag{25}$$

Instead, the error $\Delta(1_{\bar{h}})$ of this estimator for one sample item negating the hypothesis is given by

$$\Delta(1_{\bar{h}}) = |p_h - Ep_h(1_{\bar{h}})| = |p_h - [1 - Ep_h(1_h)]|. \quad (26)$$

If we have at our disposal $N$ sample items and $N \to \infty$, then the number $N_h$ of sample items confirming the hypothesis $h$ equals $p_h \cdot N$ and the number $N_{\bar{h}}$ of sample items negating the hypothesis equals $(1 - p_h) \cdot N$. Thus, the error sum $\Delta(N_h)$ of individual conclusions from all sample items confirming the hypothesis is determined by the formula (27) and the error sum $\Delta(N_{\bar{h}})$ of all sample items negating the hypothesis is determined by (28),

$$\Delta(N_h) = |p_h - Ep_h(1_h)| p_h N, \quad (27)$$

$$\Delta(N_{\bar{h}}) = |p_h - [1 - Ep_h(1_h)]|(1 - p_h)N. \quad (28)$$

The error sum $\Delta(N)$ of all individual conclusions from $N$ sample items is determined by

$$\begin{aligned}\Delta(N) = {} & |p_h - Ep_h(1_h)| p_h N \\ & + |p_h - [1 - Ep_h(1_h)]|(1 - p_h)N.\end{aligned} \quad (29)$$

From (29) we can calculate the average error $\Delta^{aver}(1)$ of probability estimation from one sample item, independently of whether the sample item confirms $(1_h)$ or negates $(1_{\bar{h}})$ the hypothesis,

$$\begin{aligned}\Delta^{aver}(1) &= \Delta(N)/N \\ &= |p_h - Ep_h(1_h)| p_h \\ &\quad + |p_h - [1 - Ep_h(1_h)]|(1 - p_h) \\ &= \Delta^{aver}(1_h) + \Delta^{aver}(1_{\bar{h}})\end{aligned} \quad (30)$$

The average error $\Delta^{aver}(1)$ of a single sample item consists of the part $\Delta^{aver}(1_h)$ representing conclusion errors from sample items $1_h$ confirming the hypothesis and the part $\Delta^{aver}(1_{\bar{h}})$ representing sample items $1_{\bar{h}}$ negating the hypothesis. The functional surface of the first part $\Delta^{aver}(1_h)$ of the whole error $\Delta^{aver}(1)$ is demonstrated in Fig. 4.

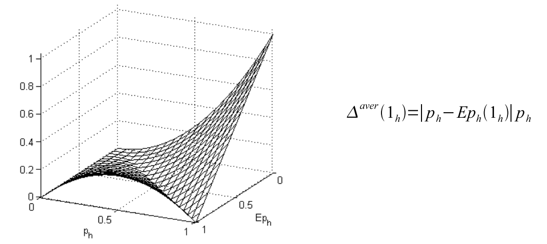


Fig. 4. Functional surface of the dependence $\Delta^{aver}(1_h) = f(p_h, Ep_h(1_h))$ of the first component of the total error $\Delta^{aver}(1)$ representing weighted estimation errors of sample items $1_h$ confirming the hypothesis $h$.

The functional surface of the second component $\Delta^{aver}(1_{\bar{h}})$ of the entire error is shown in Fig. 5.

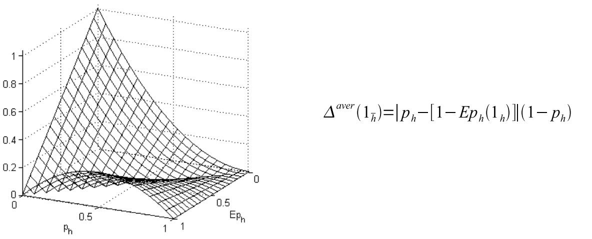


Fig. 5. Functional surface of the dependence $\Delta^{aver}(1_{\bar{h}}) = f(p_h, Ep_h(1_h))$ of the second component of the whole single-sample item error $\Delta^{aver}(1)$ of probability estimation generated by sample items $1_{\bar{h}}$, negating the hypothesis $h$.

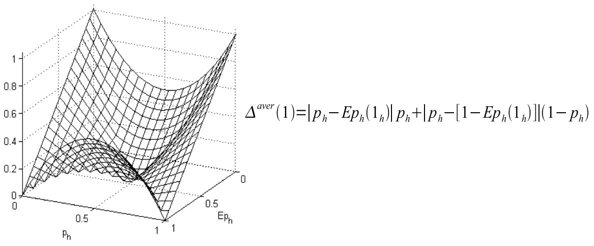


Fig. 6. Functional surface of the complete dependence $\Delta^{aver}(1) = f(p_h, Ep_h(1_h))$ of the average absolute error of one-sample item probability estimation.

Figure 6 demonstrates the functional surface of the complete dependence (30). It delivers some interesting information. The precise value $p_h$ of probability is not known and we cannot control it. However, we can control the value of $Ep_h(1_h)$ of the probability estimate from one sample item, which means that we can choose such a value of $Ep_h(1_h)$ which will minimize the risk of committing large errors of probability estimation. This risk is represented by cross-sections of the error $\Delta^{aver}(1)$—function for assumed values of $Ep_h(1_h)$. Figure 7 demonstrates the section of the functional surface from Fig. 6 for $Ep_h(1_h) = 1$. This value corresponds to the conclusion from one sample item $1_h$ made by the frequency estimator $fr_h$.

Already the visual analysis of Fig. 6 allows perception of other values of $Ep_h(1_h)$ that are better than $Ep_h(1_h) = 1$ used by the frequency estimator $fr_h = n_h/n$. An example can be $Ep_h(1_h) = 3/4$, which generates error area $A = 0.19792$ smaller than $Ep_h(1_h) = 1$, where the error area $A = 0.33333$. It allows a considerable decrease in large-error risk by the completeness estimator.

Examples presented in Figs. 7–9 show in a very clear way that assigning to the single sample item $1_h$ the radical confirmation strength $Ep_h(1_h) = 1$ by the universally used frequency estimator $fr_h = n_h/n$ is not the best idea

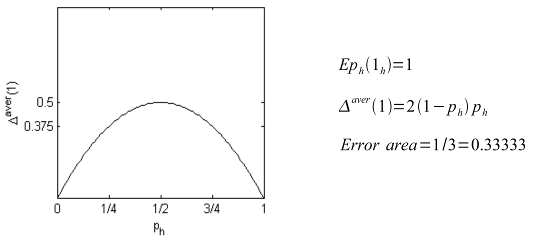


Fig. 7. Cross section of the average, single-sample item error $\Delta^{aver}(1)$ from Fig. 6 for the probability estimate $Ep_h(1_h) = 1$ that corresponds to the estimate calculated by the frequency estimator $fr_h = n_h/n$.

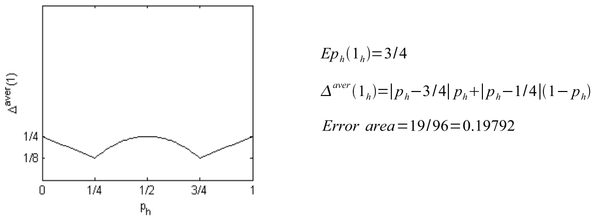


Fig. 8. Cross section of the average, one-sample item error $\Delta^{aver}(1)$ function from Fig. 6 for the estimate value $Ep_h(1_h) = 3/4$ assigned to one sample item $1_h$ confirming the hypothesis $h$.

because there are other values as, e.g., $Ep_h(1_h) = 3/4$ that considerably decrease the large-error risk of probability estimation. Further on, the optimal value of the one-sample item estimate $Ep_h(1_h)$ that minimizes the cross section area $A$ of the one-sample item error function $\Delta^{aver}(1) = f(p_h, Ep_h(1_h))$ will be derived.

The average one-sample item error $\Delta^{aver}(1)$ is expressed by

$$\begin{aligned}\Delta^{aver}(1) &= |p_h - Ep_h(1_h)|p_h \\ &\quad + |p_h - [1 - Ep_h(1_h)]|(1 - p_h).\end{aligned} \tag{31}$$

The function (31) is not continuously differentiable and changes its mathematical form in each of the four sectors $Si$ of the space of independent variables, $p_h$ and $Ep_h(1_h)$ (Fig. 10).

As can be seen in Figs. 6–9, the optimal value of the one-sample item estimate $Ep_h^{opt}(1_h)$ that minimizes the cross-section area $A$ of the error function (31) lies over the value $Ep_h(1_h) = 1/2$. Thus, the cross–section of the error function goes through the sectors $S4$, $S3$, and $S2$. Because the one-sample item error function (31) is symmetrical in relation to $p_h = 1/2$ (see Fig. 6), for calculation of the error area $A$ the formula (32)

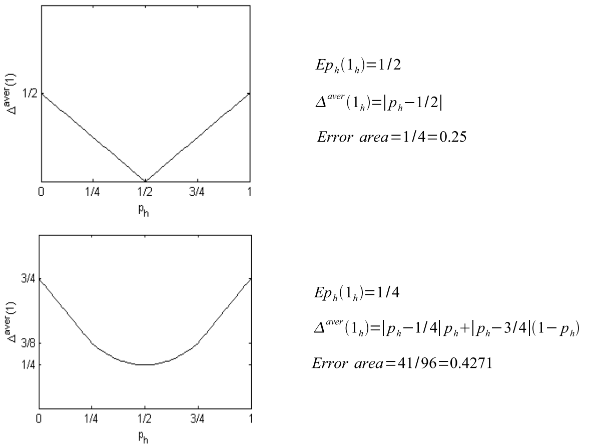


Fig. 9. Cross-sections of the average, one-sample item error function $\Delta^{aver}(1) = f(p_h, Ep_h(1_h))$ from Fig. 6 for estimate values $Ep_h(1_h) = 1/2$ and $Ep_h(1_h) = 1/4$ assigned to one sample item $1_h$ confirming the hypothesis $h$.

$$\begin{aligned}A &\\ &= 2\int_0^{1-Ep_h(1_h)} (1-2p_h)[1-Ep_h(1_h)]\mathrm{d}p_h \\ &\quad + 2\int_{1-Ep_h(1_h)}^{1/2} [2p_h(1-p_h)-[1-Ep_h(1_h)]]\mathrm{d}p_h.\end{aligned} \tag{32}$$

can be used.

After integrating the function (32), the formula for the area $A$ of the cross section of the one-sample item error function $\Delta^{aver}(1)$ is achieved,

$$\begin{aligned}A = \frac{1}{3} &- [1 - Ep_h(1_h)] + 2[1 - Ep_h(1_h)]^2 \\ &- \frac{2}{3}[1 - Ep_h(1_h)]^3.\end{aligned} \tag{33}$$

The following formula expresses the derivative $\mathrm{d}A/\mathrm{d}Ep_h(1_h)$:

$$\frac{\mathrm{d}A}{\mathrm{d}Ep_h(1_h)} = 1 - 4[1 - Ep_h(1_h)] + 2[1 - Ep_h(1_h)]^2. \tag{34}$$

After equating the derivative function (34) to zero and solving the resulting equation, the optimal value $Ep_h^{opt}(1_h)$ of the probability estimate is achieved. This value should be assigned to one sample item $1_h$ confirming the hypothesis $h$ (35),

$$Ep_h^{opt}(1_h) = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} = 0.70711. \tag{35}$$

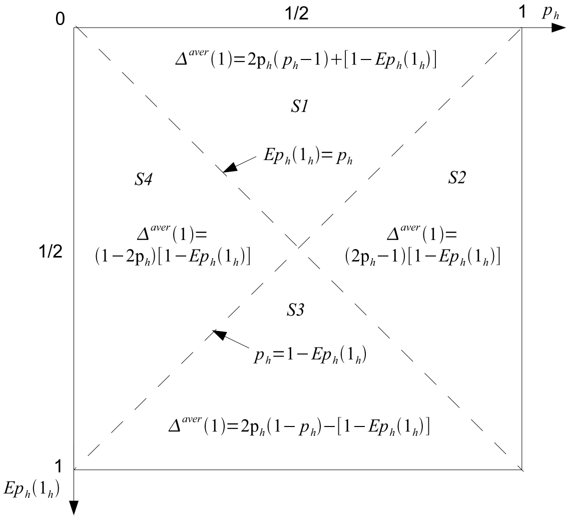

0
1/2
1
$p_h$
1/2
1
$Ep_h(1_h)$
$\Delta^{aver}(1)=2p_h(p_h-1)+[1-Ep_h(1_h)]$
S1
$Ep_h(1_h)=p_h$
S4
S2
$\Delta^{aver}(1)=$ $(1-2p_h)[1-Ep_h(1_h)]$
$\Delta^{aver}(1)=$ $(2p_h-1)[1-Ep_h(1_h)]$
S3
$p_h=1-Ep_h(1_h)$
$\Delta^{aver}(1)=2p_h(1-p_h)-[1-Ep_h(1_h)]$

Fig. 10. Four sectors $S1$–$S4$ in the space of independent variables $p_h$ and $Ep_h(1_h)$ of the one-sample item, average error function $\Delta^{aver}(1) = f(p_h, Ep_h(1_h))$.

After inserting the optimal value $Ep_h^{opt}(1_h)$ in the formula (33), the minimal error area $A_{min} = 0.19526$ is achieved. This area is a little smaller than the value $A = 0.19792$ achieved for $Ep_h(1_h) = 3/4 = 0.75$ and shown in Fig. 8. The minimal error area is shown in Fig. 11.

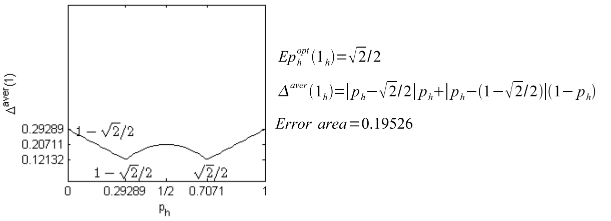


Fig. 11. Cross-section of the function $\Delta^{aver}(1)$ of the absolute, average, one-sample item error for the optimal one-sample item estimate $Ep_h^{opt}(1_h) = \sqrt{2}/2$.

In Section 2 the completeness estimator $Ep_{ha}$ of probability was proposed,

$$Ep_{ha}(1_h) = \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2(n+a)}, \tag{36}$$

where $a$ is the coefficient of the concluding caution from one sample item $1_h$ confirming the hypothesis $h$. The optimal value $Ep_h^{opt}(1_h) = \sqrt{2}/2$ can now be used for determining the optimal value of $a$. If only one sample item $1_h$ is at our disposal, then $n_h = 1$, $n_{\bar{h}} = 0$, and $n = 1$. For these values the formula (36) takes the following form:

$$Ep_{ha}(1_h) = \frac{1}{2} + \frac{1}{2(1+a)} = \frac{1}{2}\left(\frac{2+a}{1+a}\right). \tag{37}$$

Taking into account that the optimal value of $Ep_{ha}^{opt}(1_h) = \sqrt{2}/2$, we get

$$Ep_{ha}(1_h) = \frac{1}{2}\left(\frac{2+a}{1+a}\right) = \frac{\sqrt{2}}{2}. \tag{38}$$

Solving (38), the optimal value of the caution coefficient $a$ is achieved,

$$a_{opt} = \sqrt{2}.$$

Thus, the formula for the optimal probability estimator minimizing the sum of absolute estimation errors assumes the following form:

$$Ep_{h\sqrt{2}} = \frac{1}{2} + \frac{n_h - n_{\bar{h}}}{2\left(n + \sqrt{2}\right)}. \tag{39}$$

In this formula $n_h$ means the number of sample items confirming the hypothesis $h$, $n_{\bar{h}}$ the number of sample items negating the hypothesis, and $n$ the entire number of sample items ($n = n_h + n_{\bar{h}}$). The formula (39) for the optimal estimator was derived in a fully theoretical way. Thus, the following question would be very justified: "How precisely will this estimator estimate probability in practical problems?". Therefore, in the next section the results of test experiments of the completeness estimator $Ep_{h\sqrt{2}}$ will be presented.

## 4. Results of comparative experiments of probability estimation by the completeness estimator $Ep_{h\sqrt{2}}$ and the frequency estimator $fr_h = n_h/n$

Before presentation of the experimental results, the following question will be asked: "Is it at all possible to test the accuracy of probability estimation of any estimator?". In the case of a binary problem such as coin tossing, to precisely determine the probability $p_h$ of the hypothesis $h$ (head domination), an infinitely large number of experiments of coin tossing would be necessary, which is physically impossible. However, we can be supported in this task by computers. Thanks to random number generators we can get large series of 1s and 0s generated with assigned probability. Though an infinitely large series of numbers cannot be generated, computers can generate as long series as necessary to allow estimation of probability with satisfactorily large accuracy.

Computer generators have been used in random experiments by many scientist, e.g., by Larose (2010). To test and to compare the accuracy of both competitive estimators $Ep_{h\sqrt{2}}$ and $fr_h$, experiments were performed in which 1000 series with 10000 of 1s and 0s were generated with different probabilities $p_h$ of 1s: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99. Thus, the number of different $p_h$-probabilities was equal to 11. Each generated "1" should be understood as hypothesis $h$ confirmation and each generated “0” as negation of this hypothesis. Because in each experiment the $p_h$-probability was known, after generating each series of numbers, calculation of probability estimates $Ep_{h\sqrt{2}}(n)$ and $fr_h(n)$, comparison of their values with the true probability $p_h$ and calculation of absolute errors of both estimates were possible.

Figure 12 demonstrates a diagram of absolute errors the completeness and of the frequency estimator for identification of the probability $p_h = 0.5$ on the basis of small numbers of sample items $n \leq 25$. It shows a picture of only the first small part of the long series consisting of 10000 numbers.

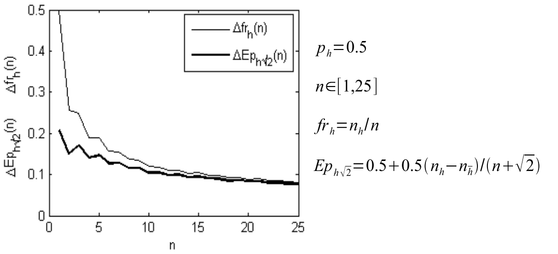


Fig. 12. Diagram of the average, absolute, estimation error $\Delta_{fr_h}(n)$ of the frequency estimator $fr_h$ and of the error $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the completeness estimator $Ep_{h\sqrt{2}}$ for small numbers of sample items $n \in [1, 25]$ calculated on the basis of 1000 experiments with 10000 numbers in each experiment. The estimated probability was $p_h = 0.5$.

Figure 12 also shows considerable differences of accuracy of both estimators, especially for very small sample item numbers $n < 10$. In particular, for $n = 1$ (single case problem), the average error of the frequency estimator equals 0.5, whereas the error of the completeness estimator equals 0.2071. The difference is greater than 100%. For the interval $n \in [6, 10]$, the average errors are $\Delta_{fr_h} = 0.1401$ and $\Delta_{Ep_{h\sqrt{2}}} = 0.1183$. For the next, greater sample item numbers, accuracy differences between both estimators disappear because both estimates converge. This phenomenon is shown in Fig. 13, where the full series of 10000 sample items is presented.

Because of the scale of Fig. 13 ($n \in [1, 10000]$), differences between both estimates for small numbers of sample items $n < 25$ are not perceptible. However, the convergence of both estimates for large $n$ is well visible. The frequency estimator identifies the probability $p_h = 0.5$ with the absolute error $\Delta_{fr_h} < 0.05$ beginning with the sample item $n_{0.05}^{fr_h} > 167$ and the completeness estimator $Ep_{h\sqrt{2}}$ beginning with the sample item

$$n_{0.05}^{Ep_{h\sqrt{2}}} > 165.$$

The estimation error $\Delta_{fr_h}(n)$ decreases below 0.01 beginning with the sample item $n_{0.01}^{fr_h} = 2784$ and the error $\Delta_{Ep_{h\sqrt{2}}}$ beginning with the sample item

$$n_{0.05}^{Ep_{h\sqrt{2}}} = 2778.$$

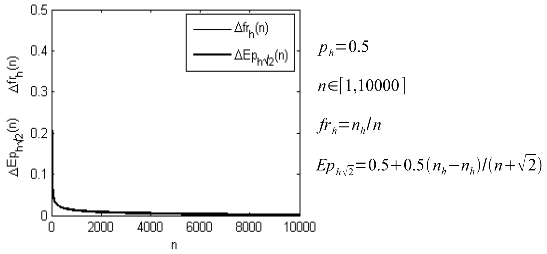


Fig. 13. Diagram of the average, absolute error $\Delta_{fr_h}(n)$ of the frequency estimator $fr_h$ and of the error $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the completeness estimator $Ep_{h\sqrt{2}}$ for full experiment series $n \in [1, 10000]$ sample items. The result is averaged for 1000 experiments.

The above results of estimation experiments were presented for the probability $p_h = 0.5$. Further on, shortly, estimation results for the following probabilities will be shown: 0.01, 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8, 0.9, 0.99.

The experiments, as expected, showed that the results for antonym probabilities, e.g., $\{p_h, 1 - p_h\}$ are almost identical and minimal differences between them are caused only by the limited number of 1000 experiments. For larger numbers of experiments, the differences would be even smaller. Figure 14 demonstrates estimation results for antonym probabilities $p_h = 0.4$ and $p_h = 0.6$.

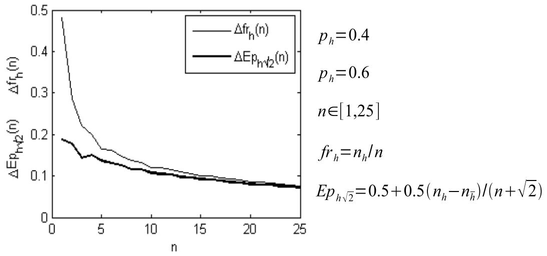


Fig. 14. Diagram of the average, absolute errors $\Delta_{fr_h}(n)$ and $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and of the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of the antonym probabilities $p_h = 0.4$ and $p_h = 0.6$ for sample item numbers $n \in [1, 25]$. The diagram shows the results averaged for 1000 experiments.

Figure 14, similarly to Fig. 12, demonstrates considerable superiority of the completeness estimator $Ep_{h\sqrt{2}}$ over the frequency one $fr_h$ in respect of accuracy. Figure 15 presents experimental results for probabilities 0.3 and 0.7.

Figure 15 also confirms the superiority of the completeness estimator $Ep_{h\sqrt{2}}$ over the frequency one $fr_h$.

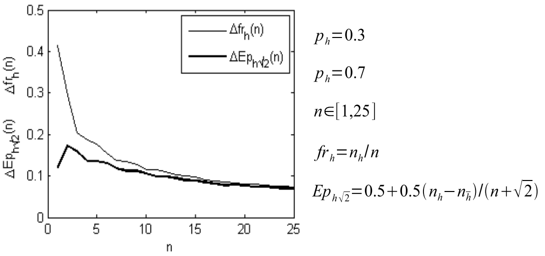


Fig. 15. Diagram of the average, absolute errors $\Delta_{fr_h}(n)$ and $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of antonym probabilities $p_h = 0.3$ and $p_h = 0.7$ for sample item numbers $n \in [1, 25]$. The results are averaged for 1000 experiments.

In comparison to the probability $p_h = 0.5$ (Fig. 12), a considerable decrease in the average estimation error for $n = 1$ to the value $\Delta_{Ep_{h\sqrt{2}}}(1) = 0.1211$ can be noticed (single case problem). Figure 16 shows results for probabilities $0.2$ and $0.8$.

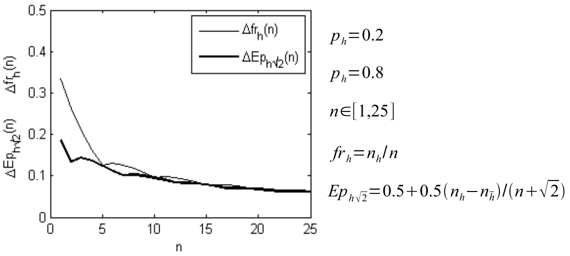


Fig. 16. Diagram of average, absolute errors $\Delta_{fr_h}(n)$ and $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and of the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of antonym probabilities $p_h = 0.2$ and $p_h = 0.8$ for sample item numbers $n \in [1, 25]$. The results are averaged for 1000 experiments.

Estimation results of probabilities $0.2$ and $0.8$ in Fig. 16 also show considerable superiority of the completeness estimator over the frequency one. Figure 17 presents experimental results for probabilities $0.1$ and $0.9$.

Figure 17 still shows certain, but not great, general superiority of the completeness estimator over the frequency one. However, for $n = 1$, the frequency estimator has a smaller error than the completeness one. The probabilities $0.1$ and $0.9$ can be called "almost certainty". Figure 18 presents experimental results for probabilities $0.01$ and $0.99$. Such probabilities can be called "practical certainty" because $p_h = 0.99$ means that in 99 of 100 cases the hypothesis is confirmed in practice. Appropriately, the probability $0.01$ means that the hypothesis in 99 of 100 cases is not confirmed in practice.

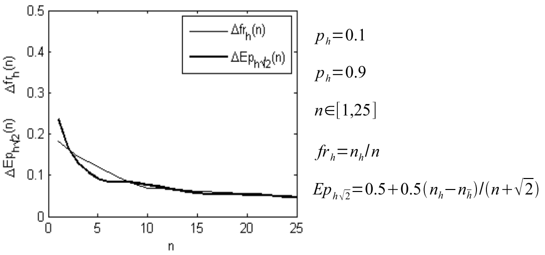


Fig. 17. Diagram of average, absolute errors $\Delta_{fr_h}(n)$ and $\Delta_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and of the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of probabilities $p_h = 0.1$ and $p_h = 0.9$ for sample item numbers $n \in [1, 25]$. The results are averaged for 1000 experiments.

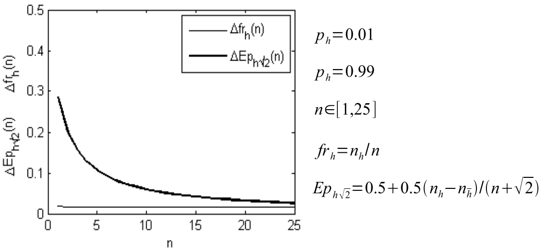
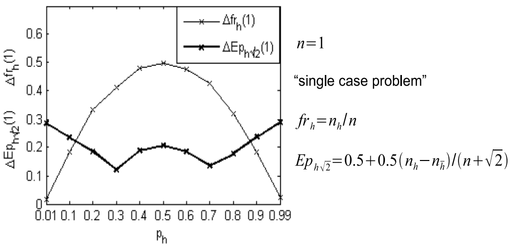


Fig. 18. Diagram of the average, absolute estimation errors $\Delta_{fr_h}(n)$ and $\Delta_{Ep_{h\sqrt{2}}}(n)$ of probabilities $p_h = 0.01$ and $p_h = 0.99$ (practical certainty) made by the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$. The results are averaged for 1000 experiments.

This time Fig. 18 shows the superiority of the frequency estimator $fr_h$ over the completeness estimator $Ep_{h\sqrt{2}}$. For all other probabilities, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, the completeness estimator was superior. Thus, it can be said that the frequency estimator is appropriate only for estimation of "easy" probabilities, i.e., such probabilities that can easily be guessed because they mean "almost certainty". Figure 19 presents the collected results of estimation errors for all 11 estimated probabilities.

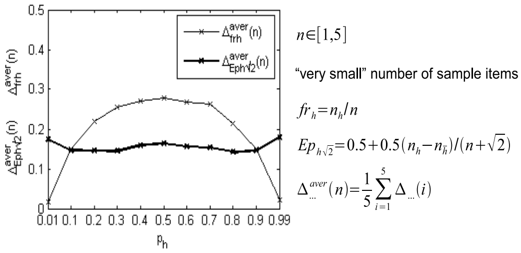
The results demonstrated in Fig. 19 are especially surprising. According to many scientists, probability estimation from one sample item makes no sense. Such an opinion has been expressed by, e.g., De Finetti (Burdzy, 2009; De Finetti, 1975). The results presented in Fig. 19 show that the completeness estimator $Ep_{h\sqrt{2}}$ allows considerable decreasing of average errors of one-sample item estimation in comparison to the frequency estimator. It should be repeated here once more that it is about average errors and not about single-case errors. In a single case the maximal error of the completeness estimator can take values in the interval $[0, 0.707]$. De Finetti was right claiming that concluding about probability from a single fact is very dangerous and should not be practised because of a great-error commitment possibility. However, sometimes (perhaps even not rarely) we are forced to derive



| $p_h$ | 0.01 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta fr_h(1)$ | 0.0178 | 0.1832 | 0.3356 | 0.4140 | 0.4820 | 0.5000 | 0.4772 | 0.4280 | 0.3212 | 0.1856 | 0.0227 |
| $\Delta Ep_{h\sqrt{2}}(1)$ | 0.2862 | 0.2360 | 0.1865 | 0.1211 | 0.1891 | 0.2071 | 0.1843 | 0.1351 | 0.1766 | 0.2372 | 0.2883 |

Fig. 19. Diagram of the average, absolute, one-sample item error $\Delta_{fr_h}(1)$ and $\Delta_{Ep_{h\sqrt{2}}}(1)$ of the frequency estimator $fr_h$ and of the completeness estimator $Ep_{h\sqrt{2}}$ from estimation of different 11 probabilities $p_h$ for $n = 1$ (single case problem). The results are averaged for 1000 experiments.

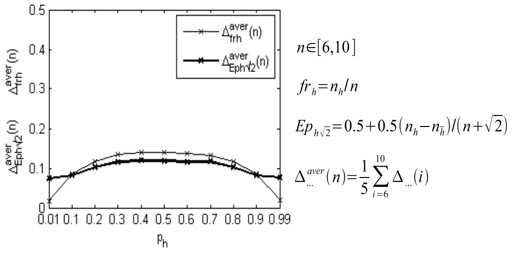
a conclusion from one fact or from one piece of data in everyday life and in professional practice. In such situations let us use the philosophy of the completeness estimator for which the maximal possible error can be equal to $0.707$ whereas in the case of the frequency estimator this error can be equal to 1, which means the $100\%$ error. Instead, the completeness estimator $Ep_{h\sqrt{2}}$ ensures a twice as small average error of concluding from one sample item than the frequency estimator $fr_h$ does (see Fig. 19). Figure 20 presents the collected errors of probability estimation from a very small number of sample items $n \in [1, 5]$.



| $p_h$ | 0.01 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta^{aver}_{frh}(n)$ | 0.0172 | 0.1495 | 0.2199 | 0.2548 | 0.2708 | 0.2770 | 0.2675 | 0.2618 | 0.2149 | 0.1490 | 0.0220 |
| $\Delta^{aver}_{Eph\sqrt{2}}(n)$ | 0.1751 | 0.1465 | 0.1452 | 0.1445 | 0.1596 | 0.1629 | 0.1552 | 0.1535 | 0.1423 | 0.1455 | 0.1783 |

Fig. 20. Diagram of the average, absolute error $\Delta^{aver}_{fr_h}(n)$ and $\Delta^{aver}_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of 11 different probabilities $p_h$ on the basis of sample item numbers $n \in [1, 5]$. The results are averaged for 1000 experiments for each $p_h$-value.

The results presented in Fig. 20 unambiguously prove that for most estimated probabilities the completeness estimator $Ep_{h\sqrt{2}}$ is considerably superior over the frequency estimator $fr_h$. The last estimator has better accuracy only for "easy" probabilities that are near 0 or 1. Instead, it makes the greatest errors in the case of the most difficult identifiable probability $p_h = 0.5$. Fig. 21 presents collected results concerning the average errors of probability estimation from a small sample item number $n \in [6, 10]$.



| $p_h$ | 0.01 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta^{aver}_{frh}(n)$ | 0.0172 | 0.0869 | 0.1158 | 0.1343 | 0.1401 | 0.1401 | 0.1374 | 0.1330 | 0.1160 | 0.0862 | 0.0192 |
| $\Delta^{aver}_{Eph\sqrt{2}}(n)$ | 0.0740 | 0.0817 | 0.1022 | 0.1161 | 0.1189 | 0.1183 | 0.1159 | 0.1165 | 0.1025 | 0.0810 | 0.0761 |

Fig. 21. Diagram of the average, absolute error $\Delta^{aver}_{fr_h}(n)$ and $\Delta^{aver}_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ for estimation of 11 different probabilities $p_h$ from sample item numbers $n \in [6, 10]$. The results are averaged for 1000 experiments for each $p_h$-value.

As Fig. 21 demonstrates, also for the sample item numbers $[6, 10]$ the completeness estimator was superior over the frequency one, apart from the probabilities that are near 0 or near 1. Figure 22 presents the collected results for sample item numbers $n \in [11, 15]$.

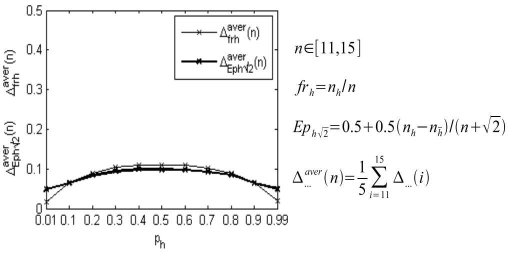
As Fig. 22 demonstrates, the completeness estimator also here is superior over the frequency one though error differences for this interval of sample item numbers are small. Figure 23 presents the collected results for sample item numbers $n \in [16, 25]$.

The differences between both competitive estimators are for this interval of sample item numbers very small because both estimators converge for larger sample item numbers.

Summing up the collected results of investigations demonstrated in Figs. 19–23, one can say without any doubt that the new completeness estimator $Ep_{h\sqrt{2}}$ is for most probabilities superior in respect of accuracy to the frequency estimator $fr_h$. In particular, it can well do things that the frequency estimator does not, i.e., it can more easily identify "difficult" probabilities that are near 0.5.

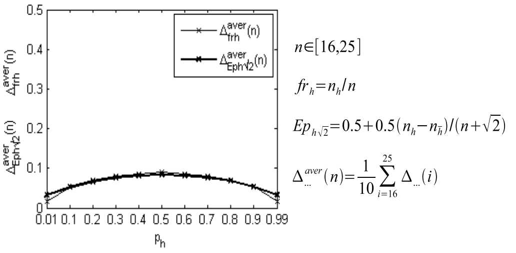
Figures 24 and 25 present diagrams of the average minimal number of sample items $n_{\ddot{0}.\ddot{0}5}$ and $n_{\ddot{0}.\ddot{0}1}$ which are necessary for estimation of various probabilities with the absolute error below $0.05$ and $0.01$.

The important knowledge they give is that the number of sample items necessary for achievement of the required accuracy of estimation strongly increases with this



| $p_h$ | 0.01 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta^{aver}_{frh}(n)$ | 0.0167 | 0.0659 | 0.0898 | 0.1043 | 0.1098 | 0.1091 | 0.1085 | 0.1019 | 0.0899 | 0.0658 | 0.0181 |
| $\Delta^{aver}_{Eph\sqrt{2}}(n)$ | 0.0475 | 0.0643 | 0.0836 | 0.0938 | 0.0989 | 0.0983 | 0.0978 | 0.0924 | 0.0844 | 0.0639 | 0.0493 |

Fig. 22. Diagram of the average, absolute error $\Delta^{aver}_{fr_h}(n)$ and $\Delta^{aver}_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ from estimation of 11 different $p_h$-probabilities for sample item numbers $n \in [11, 15]$. The results are averaged for 1000 experiments for each $p_h$-value.



| $p_h$ | 0.01 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.99 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta^{aver}_{frh}(n)$ | 0.0160 | 0.0532 | 0.0698 | 0.0807 | 0.0864 | 0.0893 | 0.0862 | 0.0802 | 0.0712 | 0.0536 | 0.0165 |
| $\Delta^{aver}_{Eph\sqrt{2}}(n)$ | 0.0317 | 0.0521 | 0.0665 | 0.0756 | 0.0806 | 0.0834 | 0.0805 | 0.0759 | 0.0686 | 0.0528 | 0.0327 |

Fig. 23. Diagram of the average, absolute error $\Delta^{aver}_{fr_h}(n)$ and $\Delta^{aver}_{Ep_{h\sqrt{2}}}(n)$ of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ from estimation of 11 different $p_h$-probabilities for sample item numbers $n \in [16, 25]$. The results are averaged for 1000 experiments for each $p_h$-value.

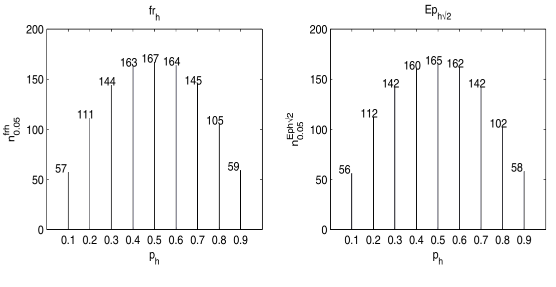


Fig. 24. Approximate, minimal sample item numbers $n^{fr_h}_{0.05}$ and $n^{Ep_{h\sqrt{2}}}_{0.05}$ necessary for identification of various probabilities $p_h$ with the use of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ with the absolute error less than $0.05$. The results are averaged for 1000 experiments for each $p_h$-value.

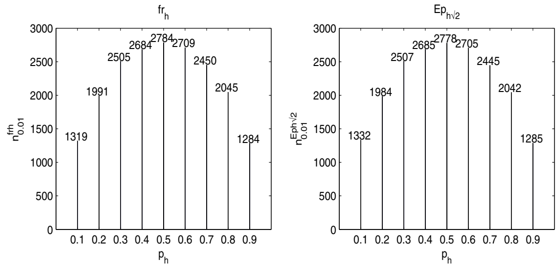


Fig. 25. Approximate, minimal sample item numbers $n^{fr_h}_{0.01}$ and $n^{Ep_{h\sqrt{2}}}_{0.01}$ necessary for identification of various probabilities $p_h$ with the use of the frequency estimator $fr_h$ and the completeness estimator $Ep_{h\sqrt{2}}$ with the absolute error less than $0.01$. The results are averaged for 1000 experiments for each $p_h$-value.

accuracy. If, e.g., the absolute error should be smaller than 0.05, then for $p_h = 0.5$ the minimal number of sample items is 165. But if the minimal error should be less than 0.01, then this number equals 2778 (for the completeness estimator).

It should be mentioned that the numbers $n^{fr_h}_{0.05}$ and $n^{fr_h}_{0.01}$, which are mean numbers identified on the basis of 1000 experiments, differ from numbers that can be calculated from the Chernoff bound (Chernoff, 1952). The Chernoff bound does not take into account calculation accuracy of probabilities estimated on the basis of a small number of sample items. And 1000 experiments are not sufficient for probability values near $0.5$. Therefore , the results presented in Figs. 24 and 25 should be treated as mean values of these particular experiments which are of only approximate character.

## 5. Comparison of the $Ep_{h\sqrt{2}}$ completeness estimator with the Cestnik–Laplace estimator

Apart from the frequency estimator also other estimators were proposed for probability estimation. It seems that the best known among them are the Cestnik and the Laplace estimator (Cestnik, 1990; 1991; Sulzmann and Furnkranz, 2009; 2010; Furnkranz and Flach, 2005). The Cestnik estimator is given by

$$p_h(n_h, n) = \frac{n_h + a}{n + a + b}. \tag{40}$$

The parameters $a$ and $b$ are degrees of freedom of the estimator and should be chosen on the basis of his/her knowledge about the investigated problem. The value $a/(a + b) = p_h(0, 0)$ means the a priori probability. The value $a + b = m$ is also problem-dependent. If little noise in the problem-data is expected, the value of $m$ should be small, if a large noise is expected, then $m$ should grow. However, in many problems, knowledge about the noise

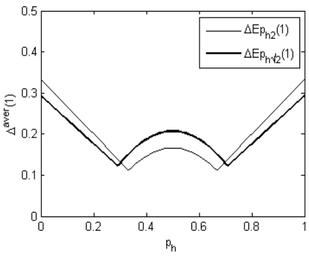


Fig. 26. Comparison of the mean absolute errors made by the two compared estimators, the $Ep_{h\sqrt{2}}$ estimator (bold line) and the Cestnik–Laplace estimator $Ep_{h2}$ (thin line) in estimation of probabilities $p_h \in [0, 1]$ on the basis of only one sample item (simple case problem).

and the a priori probability is small or even zero. Which values of $p_h(0, 0)$ and of $m$ should then be chosen? In such a situation Cichosz (2000) proposes $m = k$ and $p_h(0, 0) = 1/k$, where $k$ is number of hypotheses in the rule conclusion. In the binary case $k = 2$, $m = a + b = 2$, $p_h = 1/2$. For these values the Cestnik estimator takes the following form:

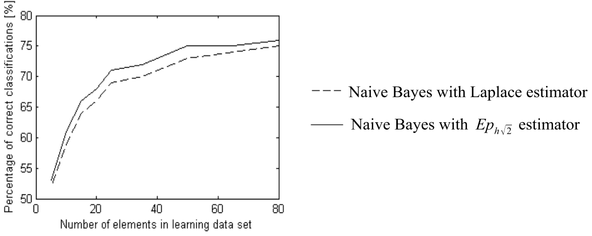
$$p_h(n_h, n) = \frac{n_h + 1}{n + 2}. \tag{41}$$

The so-achieved Cestnik estimator is identical with the classic Laplace estimator (41). This means that the Laplace estimator is a special case of the Cestnik estimator. Thus, the estimator (41) can be called the Cestnik–Laplace estimator (C–L estimator). Because the proposed $Ep_{h\sqrt{2}}$ estimator was derived with no knowledge about the a priori probability $p_h(0, 0)$ and the noise expected in the problem data, it can be compared only with the Cestnik–Laplace estimator (41) derived also with the assumption of zero knowledge concerning these parameters. Comparisons with other estimators derived under assumptions of any knowledge about the problem would be unjust and unfair. Let us notice, that the classic, frequency estimator $fr_h$ also uses no start-knowledge about the investigated problem.

Figure 26 shows a diagram of the Mean-Absolute-Errors (MAEs) of the two compared estimators at probability estimation from one sample item (single case problem).

Comparison of the $Ep_{h\sqrt{2}}$ and the C–L estimator gives the following precise values of the mean absolute error for the full probability interval $p_h \in [0, 1]$: $MAE^{aver}(1) = 0.19526$ for the $Ep_{h\sqrt{2}}$ estimator, $MAE^{aver}(1) = 0.19753$ for the C–L estimator $Ep_{h2}$.

The above results shows that the $Ep_{h\sqrt{2}}$ estimator has the general absolute $MAE^{aver}(1)$ error a little smaller than the C–L estimator in the full probability interval. However, the C–L estimator is little better for estimation of central probabilities $p_h \in [0.3, 0.7]$ and the $Ep_{h\sqrt{2}}$ estimator is a little better for estimating probabilities $p_h \in [0, 0.3]$ and $p_h \in [0.7, 1]$. Thus, the optimality interval of $Ep_{h\sqrt{2}}$ is wider. Generally, taking into account the $MAE^{aver}(1)$ criterion, the $Ep_{h\sqrt{2}}$ estimator is a little better than the Cestnik–Laplace estimator because the $Ep_{h\sqrt{2}}$ estimator was optimized on the basis of the MAE criterion.



| Number of elements in learning data set | 5 | 10 | 15 | 20 | 25 | 35 | 50 | 65 | 80 |
|---|---|---|---|---|---|---|---|---|---|
| Naive Bayes with Laplace estimator [%] | 52 | 59 | 64 | 66 | 69 | 70 | 73 | 74 | 75 |
| Naive Bayes with $Ep_{h\sqrt{2}}$ estimator [%] | 53 | 61 | 66 | 68 | 71 | 72 | 75 | 75 | 76 |

Fig. 27. Percentage of correct classifications of examples from the testing data set, depending on the number of elements in the learning data set, using naive Bayes classifiers with the Laplace estimator and with the $Ep_{h\sqrt{2}}$ estimator.

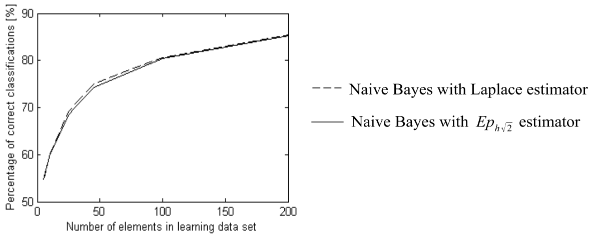
If we compare both estimators on the basis of the MSE criterion (Mean-Square-Error), then the situation is inverse. The average MSE for the full probability interval $p_h \in [0, 1]$ equals $MSE^{aver}(1) = 0.05719$ for the $Ep_{h\sqrt{2}}$ estimator, $MSE^{aver}(1) = 0.05556$ for the C–L estimator. Thus, under the MSE criterion, the C–L estimator is a little better than the $Ep_{h\sqrt{2}}$ estimator.

The performance of both estimators was also compared on two real data sets. The comparison criterion was the percentage of correct classification.

**Example 1.** The SPECT heart data set comes from UC Irvine Machine Learning Repository and describes diagnosing cardiac Single Proton Emission Computed Tomography (SPECT) images (Cios and Kurgan, 2001).

The SPECT data set has 267 instances with 23 binary attributes. Random data for the learning data set come from the SPECT.train file (80 instances) and testing data comes from the SPECT.test file (187 instances). Using two types of naive Bayes classifiers with different probability estimators (Laplace and $Ep_{h\sqrt{2}}$), the examples from data set were classified into two categories: normal and abnormal.

Figure 27 shows the results of correct classifications by naive Bayes classifiers using two estimators and different numbers of elements in the learning data set. The results are the mean of 100 experiments in each case, expressed as a percentage of correct classifications of elements from the testing data set. ♦



| Number of elements in learning data set | 5 | 10 | 15 | 25 | 25 | 30 | 45 | 75 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|---|---|
| Naive Bayes with Laplace estimator [%] | 54.43 | 60.15 | 62.98 | 66.36 | 69.00 | 70.69 | 74.98 | 78.12 | 80.67 | 85.42 |
| Naive Bayes with $Ep_{h\sqrt{2}}$ estimator [%] | 54.71 | 60.07 | 62.47 | 65.75 | 68.31 | 70.03 | 74.29 | 77.65 | 80.42 | 85.33 |

Fig. 28. Percentage of correct classifications of examples from testing data set, depending on the number of elements in the learning data set, using naive Bayes classifiers with Laplace estimator and with the $Ep_{h\sqrt{2}}$ estimator.

**Example 2.** The Balance Scale data set comes from the UC Irvine Machine Learning Repository (Siegler, 1994) and was generated to model the results of psychological experiments carried out by Siegler (1976). Using two types of naive Bayes classifiers with different probability estimators (Cestnik–Laplace and $Ep_{h\sqrt{2}}$), the examples from the data set were classified to the one of the three classes: tip of the balance scale to the right, tip to the left, or the balance scale balanced. The model consists of 625 instances and 5 attributes: class name (left, balance, right), left weight (1, 2, 3, 4, 5), left distance (1, 2, 3, 4, 5), right weight (1, 2, 3, 4, 5), and right distance (1, 2, 3, 4, 5). For example, the element of the data set "2, 5, 2, 1" (left weight = 2, left distance = 5, right weight = 2, right distance = 1) should be classified to the class "left".

Figure 28 shows the results of correct classifications by naive Bayes classifiers using two estimators and different numbers of elements in the learning data set. The results are the mean of 100 experiments in each case, expressed as a percentage of correct classifications of elements from the testing data set.

As Fig. 28 shows, both compared estimators exhibit in the case of this data set very similar performance. As can be seen in Fig. 26, the $Ep_{h\sqrt{2}}$ estimator is a little better at estimation of probabilities $p_h \in [0, 0.3]$ and $p_h \in [0.7, 1]$ and the C–L estimator at probabilities $p_h \in [0.3, 0.7]$. Thus, if approximate knowledge about the estimate probability exists, then it can be use for choosing one of the two estimators. ♦

## 6. $Ep_{h\sqrt{2}}$ completeness estimator of probability for the $n$-ary case

The number $k$ of possible hypotheses in problems is usually larger than 2 ($k > 2$). In this case the $k$-ary problem can be decomposed in $k$ binary sub-problems of the hypotheses $h_i$ and the anti-hypothesis $\bar{h}_i$ type. For each of the $k$ binary sub-problems, the binary estimate $Ep^*_{h_i\sqrt{2}}$ can be determined

$$\begin{aligned} Ep^*_{h_i\sqrt{2}} &= \frac{1}{2} + \frac{n_{h_i} - n_{\bar{h}_i}}{2\left(n+\sqrt{2}\right)} \\ &= \frac{n+\sqrt{2}+n_{h_i}-n_{\bar{h}_i}}{2(n+\sqrt{2})} \\ &= \frac{n_{h_i}+n_{\bar{h}_i}+\sqrt{2}+n_{h_i}-n_{\bar{h}_i}}{2(n+\sqrt{2})} \\ &= \frac{2n_{h_i}+\sqrt{2}}{2(n+\sqrt{2})}, \end{aligned} \qquad (42)$$

where

$$n = \sum_{j=1}^{k} n_{h_j}.$$

The sum $\sum_{j=1}^{k} Ep^*_{h_j\sqrt{2}}$ of the binary estimates calculated with (42) is not normalized in the sense of Kolmogorov's axioms (Khrennikov, 1999) and is larger than 1. Therefore, normalization is to be made. As its result the normalized $k$-nary estimate $Ep_{h_i\sqrt{2}}$ is achieved,

$$\begin{aligned} Ep_{h_i\sqrt{2}} &= \frac{Ep^*_{h_i\sqrt{2}}}{\sum\limits_{j=1}^{k} Ep^*_{h_j\sqrt{2}}} = \frac{\frac{2n_{h_i}+\sqrt{2}}{2(n+\sqrt{2})}}{\sum\limits_{j=1}^{k} \frac{2n_{h_j}+\sqrt{2}}{2(n+\sqrt{2})}} \\ &= \frac{2n_{h_i}+\sqrt{2}}{\sum\limits_{j=1}^{k}(2n_{h_j}+\sqrt{2})} = \frac{2n_{h_i}+\sqrt{2}}{2kn_{h_j}+k\sqrt{2}} \\ &= \frac{2n_{h_i}+\sqrt{2}}{2n+k\sqrt{2}} = \frac{n_{h_i}\sqrt{2}+1}{n\sqrt{2}+k}. \end{aligned} \qquad (43)$$

Attention should be paid to the fact that the optimal estimator $Ep_{h_i\sqrt{2}}$ is able to assign probability estimate also for such hypotheses that are not confirmed by any sample item ($n_{h_i} = 0$). In this case the assigned estimate equals $1/(n\sqrt{2} + k) \neq 0$. If there are no sample items at all ($n = 0$), then the estimator assigns the same probability estimate equal to $1/k$ to each hypothesis $h_i$. If there is only one sample item ($n = 1$) which confirms only one hypothesis $h_i$ (other hypotheses have no confirmations), then the estimator $Ep_{h_i\sqrt{2}}$ assigns the estimate $(\sqrt{2}+1)/(\sqrt{2}+k)$ to the hypothesis $h_i$ and to all other single hypotheses which have no confirmations the smaller estimate $1/(\sqrt{2}+k)$. In the same situation, the frequency estimator assigns the estimate 1 to the hypothesis confirmed by one sample item and 0 probability estimate to all other hypotheses without confirmation. Table 1 shows estimate values assigned to the hypothesis $h_i$ confirmed by only one sample item (single case problem, $n = 1$ and $n_{h_i} = 1$) for various number $k$ of possible hypotheses ($k = 2$: binary problem, $k = 3$: trinary problem, etc.).

Table 1. Values of the completeness probability estimate $Ep_{h_i\sqrt{2}}$ and of the frequency estimate $fr_{h_i}(1)$ assigned by both estimators to the confirmed hypothesis $h_i$ for varying number $k$ of possible hypotheses (single case problem).

| $k$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| $fr_{h_i} = \frac{n_{h_i}}{n}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $Ep_{h_i\sqrt{2}} = \frac{n_{h_i}\sqrt{2}+1}{n\sqrt{2}+k}$ | 0.707 | 0.547 | 0.446 | 0.376 | 0.326 | 0.287 | 0.258 | 0.232 | 0.211 |

Further on, we will present an example of application of the completeness estimator $Ep_{h_i\sqrt{2}}$ for probability estimation of the travel time $T_{AB}$ with a car between two places $A$ and $B$ in Szczecin (Poland), the distance between which equals 17 km. Let us assume that only 10 measurements of the travel time $T_{AB}$ [min] are at our disposal ($n = 10$). There is no confirmation ($n_{h_1} = 0$ of the travel time $19 < T_{AB}$[min] and only one confirmation ($n_{h_2} = 1$) of the travel time $19 < T_{AB} \leq 20$ [min]. There are three confirmations ($n_{h_3} = 3$) of the travel time $20 < T_{AB} \leq 21$ [min], four confirmations ($n_{h_4} = 4$) of the time $21 < T_{AB} \leq 22$ [min], two confirmations ($n_{h_5} = 2$) of the time $22 < T_{AB} \leq 23$ [min], and no confirmations ($n_{h_6} = 0$) of the time $T_{AB} > 23$ [min]. The histogram of probability of the travel time calculated on the basis of the frequency estimator and of the optimal completeness estimator $Ep_{h_i\sqrt{2}}$ should be determined. Both histograms are for comparison presented in Fig. 29.

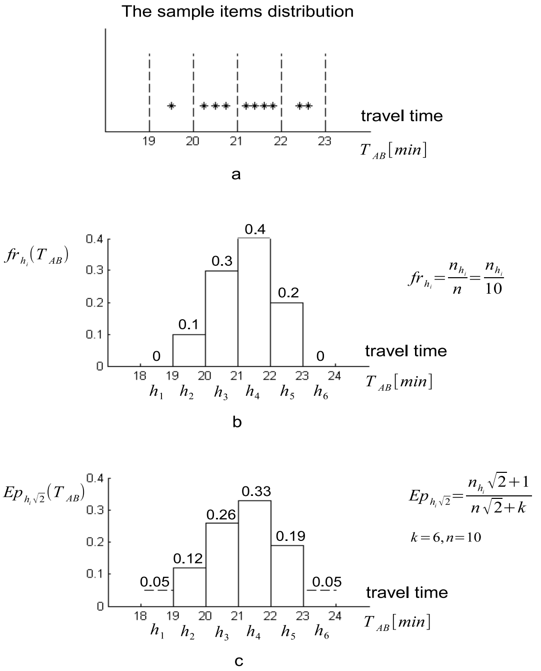


Fig. 29. Distribution of measurement sample items of the travel time $T_{AB}$ [min] between points $A$ and $B$, histogram of travel time probability determined with the frequency estimator $fr_{h_i} = n_{h_i}/n$ (b), histogram determined on the basis of the optimal completeness estimator $Ep_{h_i\sqrt{2}}$ (c).

As Fig. 29 shows, the optimal estimator $Ep_{h_i\sqrt{2}}$ does not assign probability estimates to particular hypotheses $h_i$ concerning the travel time $T_{AB}$ in such a drastic way as the frequency estimator $fr_{h_i}$ does. Probability is distributed between particular hypotheses concerning the travel time more evenly than in the case of the frequency estimator. In particular, the frequency estimator does not assign any, even minimal, probability to the hypotheses which are not confirmed by sample items ($h_1$ and $h_6$). Such assignment is not credible because next $T_{AB}$: time measurements (sample items) can be larger than 23 min (traffic jams can occur) or smaller than 19 min. As the theoretical proofs and experiments described in the paper show, the completeness estimator $Ep_{h_i\sqrt{2}}$ is optimal in the sense of the absolute error-sum and the frequency estimator $fr_{h_i}$ is not. Therefore, probability estimates delivered by the estimator $Ep_{h_i\sqrt{2}}$ are more precise and credible than estimates delivered by the frequency estimator $fr_{h_i}$. A very similar situation as in construction of probability histograms for one single variable exists in the case of many variables when the probability of rules are evaluated for rules achieved with decision trees, rough sets theory, fuzzy sets theory, data mining methods and other methods of artificial intelligence.

## 7. Conclusions

The paper presents a new, optimal probability estimator $Ep_{h\sqrt{2}}$ that minimizes the sum of absolute estimation errors. The author of the estimator idea and of the theoretical proof of its optimality is Andrzej Piegat. The theory of the estimator needed experimental verification. Computer programs towards this verification and the experiments were realized by Marek Landowski. The experiments were of comparative character. Their aim was to determine the accuracy of the new completeness estimator and to compare it with that of the universally used frequency estimator $fr_h = n_h/n$ and with the $m$-estimator of Cestnik. The experiments, similarly to the theoretical derivation, showed a considerably greater accuracy of the completeness estimator than that of the frequency one in the case of small sample item numbers $n < 25$ and the same accuracy for larger sample item numbers $n > 25$. Because in many real problems we have to deal with data insufficiency, the new completeness estimator seems very useful and practice-oriented. Comparison of the $Ep_{h\sqrt{2}}$ estimator with the Cestnik–Laplace estimator showed a little better performance of the former. However, it depends on the value of an estimated probability.

The $Ep_{h\sqrt{2}}$ estimator parameter $a$ is not assumed by anybody in the subjective, a priori way. It was derived on the basis of the optimization criterion of the sum of absolute estimation errors. The paper also presents a solution for the single-case problem.

## References

Ben-Haim, Y. (2006). *Info-gap Decision Theory*, Elsevier, Oxford/Amsterdam.

Burdzy, K. (2009). *The Search for Certainty. On the Clash of Science and Philosophy of Probability*, World Scientific, Singapore.

Burdzy, K. (2011a). Blog on the book *The Search for Certainty. On the Clash of Science and Philosophy of Probability*, http://search4certainty.blogspot.com/.

Burdzy, K. (2011b). Philosophy of probability, Website, http://www.math.washington.edu/∼burdzy/philosophy/.

Carnap, R. (1952). *Logical Foundations of Probability*, University Press, Chicago, IL.

Cestnik, B. (1990). Estimating probabilities: A crucial task in machine learning, *in* L. Aiello (Ed.), *ECAI'90*, Pitman, London, pp. 147–149.

Cestnik, B. (1991). *Estimating Probabilities in Machine Learning*, Ph.D. thesis, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana.

Chernoff, H. (1952). A measure of asymptotic efficiency for test of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics* **23**(4): 493–507.

Cichosz, P. (2000). *Learning Systems*, Wydawnictwa Naukowo-Techniczne, Warsaw, (in Polish).

Cios, K. and Kurgan, L. (2001). SPECT heart data set, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/spect+heart.

De Finetti, B. (1975). *Theory of Probability: A Critical Introductory Treatment*, Willey, London.

Dubois, D. and Prade, H. (1988). *Possibility Theory*, Plenum Press, New York/NY, London.

Furnkranz, J. and Flach, P.A. (2005). Roc'n'rule learning: Towards a better understanding of covering algorithms, *Machine Learning* **58**(1): 39–77.

Hajek, A. (2010). Interpretations of probability, *in* E.N. Zalta, (Ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/probability-interpret/.

Khrennikov, A. (1999). *Interpretations of Probability*, Brill Academic Pub., Utrecht/ Boston, MA.

Klirr, G.J. and Yuan, B. (1996). *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems. Selected Papers by Lotfi Zadeh*, World Scientific, Singapore.

Laplace, P.S. (1814, English edition 1951). *A Philosophical Essay on Probabilities*, Dover Publication, New York/NY.

Larose, D.T. (2010). *Discovering Statistics*, W.H. Freeman and Company, New York, NY.

Piegat, A. (2011a). Uncertainty of probability, *in* K.T. Atanassov, M. Baczyński, J. Drewniak, J. Kacprzyk, M. Krawczak, E. Schmidt, M. Wygralak and S. Zadrożny (Eds.) *Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics, Vol. I: Foundations*, IBS PAN, Warsaw, pp. 159–173.

Piegat, A. (2011b). Basic lecture on completeness interpretation of probability, Website, http://kmsiims.wi.zut.edu.pl/pobierz-pliki/cat view/47-publikacje.

Polkowski, L. (2002). *Rough Sets*, Physica-Verlag, Heidelberg/New York, NY.

Popper, K.R. (1957). The propensity interpretation of the calculus of probability and the quantum theory, *in* S. Korner (Ed.), *Observation and Interpretation: A Symposium of Philosophers and Physicists*, Butterworth Scientific Publications, London, pp. 65–70.

Rocchi, P. (2003). *The Structural Theory of Probability: New Ideas from Computer Science on the Ancient Problem of Probability Interpretation*, Kluwer Academic/Plenum Publishers, New York, NY.

Rokach, L. and Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*, Machine Perception and Artificial Intelligence, Vol. 69, World Scientific Publishing, Singapore.

Shafer, G. (1976). *A Mathematical Theory of Evidence*, Princetown University Press, Princetown, NJ .

Siegler, R.S. (1976). Three aspects of cognitive development, *Cognitive Psychology* **8**(4): 481–520.

Siegler, R.S. (1994). Balance scale weight & distance database, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/balance+scale.

Sulzmann, J.N. and Furnkranz, J. (2009). An empirical comparison of probability estimation techniques for probabilistic rules, *in* J. Gama, J. Santos Costa, A.M. Jorge and P. Brazdil (Eds.), *Proceedings of the 12th International Conference on Discovery Science (DS-09)*, Springer-Verlag, Heidelberg/New York, NY, pp. 317–331.

Sulzmann, J.N. and Furnkranz, J. (2010). Probability estimation and aggregation for rule learning, *Technical Report TUD-KE-201-03*, Knowledge Engineering Group, TU Darmstadt, Darmstadt.

von Mises, R. (1957). *Probability, Statistics and the Truth*, Macmillan, Dover/New York, NY.

Witten, I.H. and Frank, E. (2005). *Data Mining*, Elsevier, Amsterdam.

Zadeh, L.A. (1965). Fuzzy sets, *Information and Control* **8**(3): 338–353.

Ziarko, W. (1999). Decision making with probabilistic decision tables, *in* N. Zhong (Ed.), *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, Proceedings of the 7th International Workshop, RSFDGrC99, Yamaguchi, Japan*, Springer–Verlag, Berlin/Heidelberg, New York, NY, pp. 463–471.

**Andrzej Piegat** received his Ph.D. degree in 1979 in modeling and control of production systems from the Technical University of Szczecin, the D.Sc. degree in control of underwater vehicles from Rostock University in 1998, and the professorial title in 2001. At present he is a professor at the West Pomeranian University of Technology. His current research is focused on uncertainty theory, fuzzy logic, computing with words and info-gap theory.

**Marek Landowski** received the M.Sc. degree in mathematics from Szczecin University in 2002, the M.Sc. in computer science from the Szczecin University of Technology in 2004, and the Ph.D. degree in 2009 (identification of probabilistic models of human perception) from the West Pomeranian University of Technology. Currently he is an assistant professor at Szczecin Maritime University. At present his research interests are focused on data mining, decision trees and info-gap theory.

Received: 25 May 2011