Dear Reader,

Data is one of the resources on which modern civilization is built. Like most resources, it requires processing before it becomes fully useful. The intellectual refinery that humanity has developed to deal with the abundant deposits of information is broadly termed statistics. Statistics also serves as a common denominator for the articles published in this edition of *Delta*, the Polish popular science monthly that you are holding in your hands. In almost 50 years of its history it has been striving to bring the subject areas that it covers – mathematics, informatics, physics and astronomy – closer to its readers. In a similar way, statistics brings the knowledge hidden in data closer to researchers and, in turn, the society.

The creation of this issue is correlated with the 34th conference "European Meeting of Statisticians" held in Warsaw from July 3rd to 7th, 2023. The local organization of the conference is by the Polish Mathematical Society, Warsaw University of Technology, and the University of Warsaw; the latter is also a publisher of *Delta*. The participants have been given copies of the English version of this edition of *Delta*. It is true that participants of this conference hardly need an additional education in statistics, but we believe that everyone will find some intellectual stimulation on these pages, regardless of their scientific background and experience.

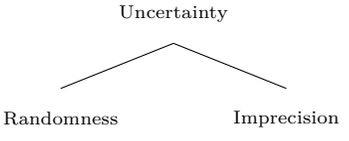
We wish you pleasant reading and, if you are a participant of the conference, a memorable event.

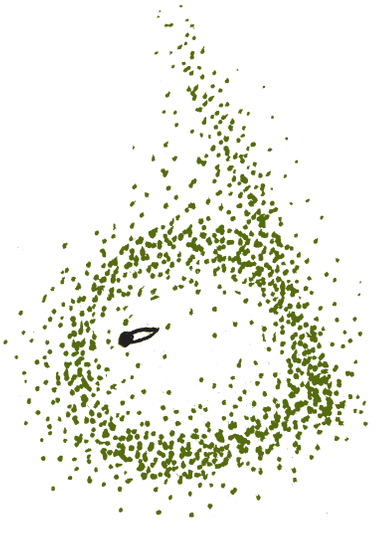
Editorial Board

Statistics with imprecise data

*Przemysław GRZEGORZEWSKI**

* Faculty of Mathematics and Information Science, Warsaw University of Technology





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

Statistics might be perceived as an art of making decisions in the presence of uncertainty. It delivers tools for describing and explaining reality as well as for making predictions and verifying hypotheses. For a long time, uncertainty has been identified with randomness, and consequently, probability has been perceived as the only well-grounded theory of uncertainty. However, during the last fifty years, several approaches extending or orthogonal to the classical probability theory have appeared. A common feature of these new approaches is an attempt to soften the classical methods so that they can more easily adapt to the factual nature of the data available and deal with other types of uncertainty, such as imprecision.

It is important to remember that imprecision as a concept itself is not entirely unambiguous. Quite often the results of an experiment are imprecise due to inaccuracy of the measuring apparatus or errors made by the persons making the measurements. Sometimes the desired measurement is so difficult that its result, as a rule, should be treated as highly uncertain. It may also happen that the exact value of a variable is intentionally hidden for some confidentiality reasons. In all these situations data are often recorded as set-valued objects (e.g. as intervals) containing the exact but unknown values so a set-valued observation A delivers incomplete information about the point quantity x: we know only that A contains x but the true value of x remains unknown. Hence A represents the **epistemic** state of the subject. But there are also situations when the experimental data appear as essentially imprecise. A typical case is the analysis of perceptions collected from a human when there is no objective value behind (like the taste or mood). Another example refers to objects or phenomena with an intrinsically gradual representation subject to variability in nature, with fuzzy or changing boundaries, flexible time intervals or rating scales, etc. Each such observation represents an objective entity, even if it is vague, and hence corresponds to **ontic** imprecision.

A convenient method of mathematical modeling of imprecision was indicated by Lotfi A. Zadeh (1921–2017) who introduced **fuzzy set theory** as an extension of the classical set theory. Zadeh, one of the most outstanding thinkers of the current time, realized that although we are used to dividing everything into "yes" and "no" or to black and white, the entire world is in shades of grey. His famous statement that *everything is a matter of degree* became the main idea behind fuzzy logic and its impressive applications. It is worth noting that fuzzy logic is actually not a "fuzzy" logic, but a logic that describes and tames imprecision.

A fundamental Zadeh's concept is a **fuzzy set**. Let $\mathbb{U}$ be a universe of discourse. A **fuzzy set** A in $\mathbb{U}$ is identified with a mapping, called a **membership**

Note that in the set theory functions $f\colon X \to Y$ are usually defined as sets; more precisely, subsets of $X \times Y$ such that for all $x \in X$ there exists an unique $y \in Y$ satisfying $(x, y) \in f$. In this sense $f(x)$ is only a notational convention to denote y for which $(x, y) \in f$.

function $A : \mathbb{U} \to [0, 1]$, which assigns to each object $x \in \mathbb{U}$ a real number in the interval $[0, 1]$, so that $A(x)$ represents the degree of membership of x into A. Thus a fuzzy set A may be perceived as a (standard) subset of $\mathbb{U} \times [0, 1]$

$$A = \{(x, A(x)) : x \in \mathbb{U}, A(x) \in [0, 1]\}.$$

The interpretation of the membership function is natural: if $A(x) = 1$ then we are sure that element x belongs to A, while $A(x) = 0$ means that x does not belong to A. In all other cases, i.e. if $A(x) \in (0, 1)$, we have a partial membership (belongingness) to A. It means that if $A(x)$ is close to 1 then the degree of membership of x in A is high, while if $A(x)$ is close to 0 then the degree of membership of x in A is low. If $A(x) \in \{0, 1\}$ for all $x \in \mathbb{U}$ then A is a set in the classical meaning (each "usual" set is a fuzzy set whose membership function is its characteristic function).

Another important notion connected with a fuzzy set is the so-called **α-cut**. For each $\alpha \in [0, 1]$ the α-cut of a fuzzy set A, denoted as A_α, is given by

$$A_\alpha = \begin{cases} \{x \in \mathbb{U} : A(x) \geqslant \alpha\} & \text{if } \alpha \in (0, 1], \\ \mathrm{cl}\{x \in \mathbb{U} : A(x) > 0\} & \text{if } \alpha = 0, \end{cases}$$

where cl stands for the closure (for now on we assume that $\mathbb{U}$ is equipped with such operation). In other words, the α-cut is a "usual" subset of $\mathbb{U}$ whose degree of belonging to A is not less than α. It can be shown that every fuzzy set is completely characterized by a family of all its α-cuts $\{A_\alpha\}_{\alpha \in [0,1]}$. Two α-cuts are of special interest: A_1 known as the **core**, which contains all values which are fully compatible with the concept described by A and A_0 called the **support**, which are compatible to some extent with the concept modeled by A.

An important subfamily of fuzzy sets are fuzzy numbers. We say that A is a **fuzzy number** if $A : \mathbb{R} \to [0, 1]$ such that its α-cuts for each $\alpha \in [0, 1]$ are nonempty closed intervals. An example of a fuzzy number is shown in Fig. 1.

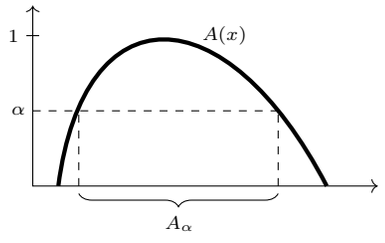


Fig. 1. A membership function $A(x)$ of a fuzzy number A.

Example. Gamonedo cheese is a kind of blue cheese produced in Asturias (northern Spain). It experiences a smoking process and later on is left to settle in natural caves or a dry place. To maintain the quality of the cheese, experts (tasters) express their subjective perceptions about different characteristics of the cheese, such as visual parameters (shape, rind, appearance), texture parameters (hardness and crumbliness), olfactory-gustatory parameters (smell intensity, smell quality, flavor intensity, flavor quality, and aftertaste) and an overall impression of the cheese. Recently tasters were asked to express their subjective perceptions about the quality of the Gamonedo cheese by using fuzzy numbers. This type of fuzzy number is the most commonly used for fuzzy descriptions both because is easy to understand by the tasters and simple in further processing. Valuation of the different features of each cheese is made over a graduate scale ranging from 0% (for lowest quality) to 100% (for highest quality). The 0-level is the set of values considered by a tester as compatible with his opinion to some extent, i.e., he thinks it is not possible that the quality is out of this set. The 1-level is the set of values considered as fully compatible with his opinion. For example, Fig. 2 illustrates a situation of a tester who believes that a given cheese meets the quality requirements in terms of the examined feature in 70–80%. At the same time, he is undoubtedly convinced that the quality requirements are satisfied with not lower than 50% but not higher than 90%.

Ramos-Guajardo A.B., et al., *Applying statistical methods with imprecise data to quality control in cheese manufacturing*, In: Grzegorzewski P., et al. (Eds.), Soft Modeling in Industrial Manufacturing, Springer 2019, pp. 127–147.

Solution to Problem M 1750.
Consider any arrangement of numbers. Note that the numbers from 1 to 2022 represent at most 2022 rows and 2022 columns. Therefore, there exists a row and a column that contain only numbers greater than 2022. The product of any two of these numbers is at least $2023 \cdot 2024$, which is greater than any number on the board. This means that there is no arrangement of numbers satisfying the conditions stated in the problem.

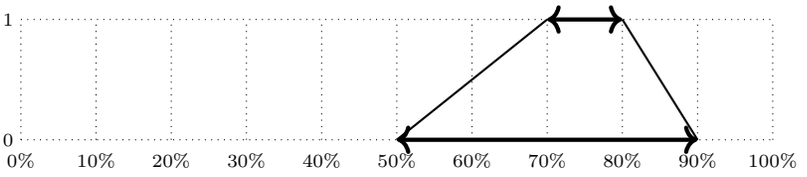
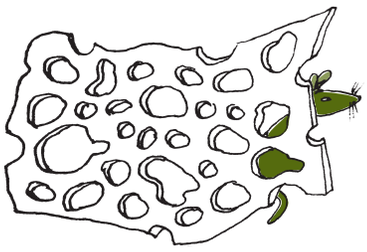


Fig. 2. An exemplary opinion of a taster expressed by a trapezoidal fuzzy set.

As is seen in the figure, both 0-level and 1-level are linearly interpolated to get the so-called trapezoidal fuzzy set used later to represent this tester's personal



valuation. A sample of tester's opinions modeled with trapezoidal fuzzy sets is given in Fig. 3.

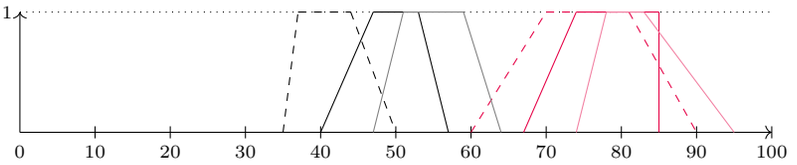


Fig. 3. A sample of tester's opinions modeled with trapezoidal fuzzy sets.

More formally, we say, that A is a **trapezoidal fuzzy number** if its membership function is given by

$$A(x) = \begin{cases} \frac{x-a_1}{a_2-a_1} & \text{if } a_1 \leqslant x < a_2, \\ 1 & \text{if } a_2 \leqslant x \leqslant a_3, \\ \frac{a_4-x}{a_4-a_3} & \text{if } a_3 < x \leqslant a_4, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $a_1, a_2, a_3, a_4 \in \mathbb{R}$ such that $a_1 \leqslant a_2 \leqslant a_3 \leqslant a_4$. Thus, since a trapezoidal fuzzy number (1) is characterized completely by four real numbers, it is often denoted as $A = (a_1, a_2, a_3, a_4)_T$.

It is worth noting that a sum of trapezoidal fuzzy numbers is also a trapezoidal fuzzy number, i.e., if $A = (a_1, a_2, a_3, a_4)_T$ and $B = (b_1, b_2, b_3, b_4)_T$ then

$$A + B = \\ = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4)_T.$$

Before we take the next step we have to define some basic operations on fuzzy numbers. Although one can introduce these operations directly on membership functions it seems that it is easier to do this equivalently as α-cut-wise operations on intervals. In particular, the sum of two fuzzy numbers A and B is given by the Minkowski addition of the corresponding α-cuts (see Fig. 4), i.e. for all $\alpha \in [0, 1]$

$$(A + B)_\alpha = \big[\inf A_\alpha + \inf B_\alpha, \sup A_\alpha + \sup B_\alpha\big]. \tag{2}$$

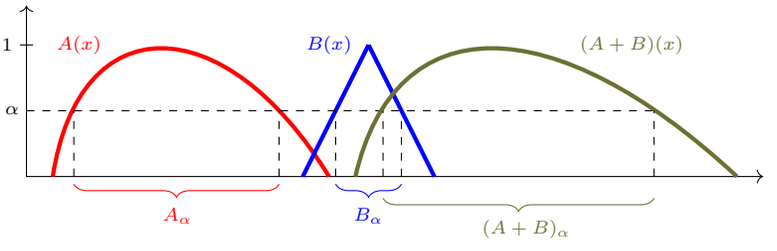


Fig. 4. Addition of fuzzy numbers A and B.

The product of a trapezoidal fuzzy number $A = (a_1, a_2, a_3, a_4)_T$ by a scalar θ is a trapezoidal fuzzy number, i.e.

$$\theta \cdot A = \begin{cases} (\theta a_1, \theta a_2, \theta a_3, \theta a_4)_T & \text{if } \theta \geqslant 0, \\ (\theta a_4, \theta a_3, \theta a_2, \theta a_1)_T & \text{if } \theta < 0. \end{cases}$$

Similarly, the product of a fuzzy number A by a scalar $\theta \in \mathbb{R}$ is defined by the Minkowski scalar product for intervals (see Fig. 5), i.e. for all $\alpha \in [0, 1]$

$$(\theta \cdot A)_\alpha = \big[\min\{\theta \inf A_\alpha, \theta \sup A_\alpha\}, \max\{\theta \inf A_\alpha, \theta \sup A_\alpha\}\big]. \tag{3}$$

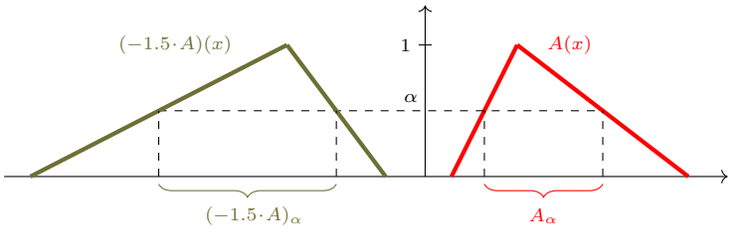


Fig. 5. The product of a fuzzy number A by a scalar.

Unfortunately, in general, $A + (-1 \cdot A) \neq \mathbb{1}_{\{0\}}$ (see Fig. 6). Consequently, the Minkowski-based difference does not satisfy, in general, the addition/subtraction property that $(A + (-1 \cdot B)) + B = A$.

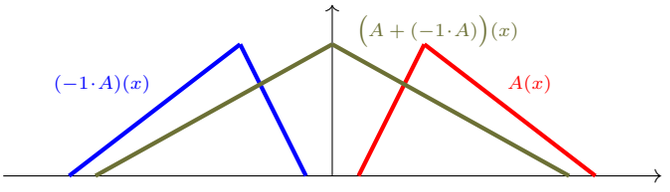


Fig. 6. Problems with subtraction of fuzzy numbers.

Solution to Problem M 1752.
Let us observe that for any $M, a, b \in [0,1]$ such that $M \geqslant a, b$, the following inequality holds:

$$(M-a)(M-b)(1-bM) \geqslant 0,$$

After some transformations, we obtain:

$$(1-bM+b^2)(1-aM+M^2) \geqslant 1-ab+b^2.$$

To prove the inequality for $n=2$, it is sufficient to take $a=b=x_1$ and $M=x_2$ in the above inequality.
Assume that the inequality holds for some n; we will deduce its validity for $n+1$. Without loss of generality, assume that $x_{n+1} = \max\{x_1, \ldots, x_{n+1}\}$. Using the above inequality with $b = x_n$, $M = x_{n+1}$, and $x_1 = a$, we obtain:

$$(1-x_n x_{n+1}+x_n^2)(1-x_{n+1}x_1+x_{n+1}^2) \geqslant \\ \geqslant (1-x_n x_1+x_n^2).$$

Therefore,

$$\prod_{\text{cycle}}^{n+1}(1-x_i x_{i+1}+x_i^2) \geqslant \\ \geqslant \prod_{\text{cycle}}^{n}(1-x_i x_{i+1}+x_i^2) \geqslant 1,$$

where "cycle" denotes the cyclic product. We complete the proof by invoking the principle of mathematical induction.

To overcome some of the problems associated with the lack of a satisfying difference, especially in constructing tools for statistical reasoning based on fuzzy observations, an alternative approach utilizing distances is often considered. Let us define the following distance between two fuzzy numbers A and B

$$D(A,B) = \sqrt{\int_0^1 \left[(\inf A_\alpha - \inf B_\alpha)^2 + (\sup A_\alpha - \sup B_\alpha)^2\right] d\alpha}. \tag{4}$$

Indeed, (4) defines a *metric* (in the sense explained in details in e.g. Jarosław Górnicki's article from *Delta* 5/2021). It is clear that $D(A,B) \geqslant 0$ and $D(A,B) = 0$ if and only if $A = B$. Proving that $D(A,B) + D(B,C) \geqslant D(A,C)$ (triangle inequality) is slightly less trivial and we leave is as an exercise to the reader.

Suppose, we observe independently two fuzzy random samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ drawn from two populations (each x_i and y_i is a fuzzy number). We want to check if there is a significant difference between these two populations. To this end we measure the distance (4) between the arithmetic means of these samples. Note that we already know how to compute an arithmetic mean of fuzzy numbers (which is itself a fuzzy number) as we have tools of adding them together and multiplying by real numbers. But is a specific distance between means, like 3.14, large or small? This is where the statistics come into the picture.

In statistical jargon our goal is to verify the null hypothesis H_0 that both samples come from the same distribution, against the alternative hypothesis that the population distributions differ. If the null hypothesis holds we expect that both sample means would not differ too much. On the other hand, a significant difference between the two sample means may indicate that the samples under study come from different distributions.

To decide whether the measurement is large enough to conclude as significant statisticians often use the notion of *p-value*. In our case it is the probability, under the assumption that the null hypothesis is true, of getting at least as great distance between means as the distance observed. Intuitively, if this probability is low, we have a good reason to reject the null hypothesis. The problem is that in this case we cannot compute it exactly as the null hypothesis merely says that the two populations can be treated as one but it does not give us a specific description of this population! We need to resort to another clever idea.

Let $\boldsymbol{v}$ be the concatenation of the two samples, i.e. $v_i = x_i$ if $1 \leqslant i \leqslant n$ and $v_i = y_{i-n}$ if $n+1 \leqslant i \leqslant N$, where $N = n+m$. Now, let $\boldsymbol{v}^*$ denote a permutation of the initial dataset $\boldsymbol{v}$. Then the first n elements of $\boldsymbol{v}^*$ are assigned to the first sample $\boldsymbol{x}^*$ and the remaining m elements to $\boldsymbol{y}^*$. In other words, it works like a random assignment of elements into two samples of the size n and m, respectively. Each permutation corresponds to some relabeling of the combined dataset $\boldsymbol{v}$. Please note that if H_0 holds, i.e. both samples come from the same distribution, then we are completely free to exchange the labels x or y attributed to particular observations – this will not change the randomness behind them. As a consequence we can *estimate* the true p-value from the data by taking a fraction of all possible permutations $\boldsymbol{v}^*$ that yield a larger distance between means of $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ that the one observed.

Formally this can be expressed as

$$\text{p-value} = \frac{1}{N!}\sum_{\boldsymbol{v}^*} \mathbb{1}(T(\boldsymbol{v}^*) \geqslant t_0), \tag{5}$$

where the sum ranges over all possible permutations $\boldsymbol{v}^*$ of $\boldsymbol{v}$, $T(\boldsymbol{v}^*)$ is the distance between means of $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ and t_0 is the observed distance between

Equation (5) can be treated as a definition of the *true* p-value, conditioning on the event that our samples sum up (as sets) to $\boldsymbol{v}$. This approach is somewhat standard in designing so called *nonparametric tests*, like *runs test*, described in details in an article of the same name in *Delta* 9/2017.

means of $\boldsymbol{x}$ and $\boldsymbol{y}$. The value of $\mathbb{1}(\texttt{condition})$ is 1 if condition is met and 0 otherwise.

Formula (5) can be further simplified, using the fact that the permutations can be split into groups of size $n!m!$ each, which give the same means of $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ (those groups are permutations that can be obtained by each other by permuting first n and last m observations). Even in this case number of summands (which becomes $\binom{N}{n}$) grows exponentially with N (given that n/N is kept at fixed level).

Therefore, instead of considering all possible permutations we consider an approximate distribution obtained by drawing randomly a large number of samples (permutations) with replacement.

Let $\boldsymbol{v}_1^*, \boldsymbol{v}_2^*, \ldots, \boldsymbol{v}_K^*$ be some random permutations of $\boldsymbol{v}$ (where K is usually not smaller than 1000). Then the approximate p-value of our test is given by

$$\text{p-value} \simeq \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(T(\boldsymbol{v}_k^*) \geqslant t_0). \tag{6}$$

Example. Now we utilize some data given in Ramos-Guajardo et al. (2019) to compare the opinions of the two experts about the overall impression of the Gamonedo cheese. The trapezoidal fuzzy sets corresponding to their opinions are gathered in Table 1. There we have two observations of independent samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_m)$ of sizes $n = 40$ and $m = 38$, respectively. Numbers in parentheses correspond to the notation used to describe trapezoidal fuzzy numbers, e.g. $x_1 = (65, 75, 85, 85)_T$, $y_1 = (50, 50, 63, 75)_T$, etc. Our problem is to check whether there is a general agreement between these two experts. To reach the goal we verify the following null hypothesis H_0 stating there is no significant difference between experts' opinions, against that their opinions on the cheese quality differ.

Table 1. Opinions of two experts concerning the overall impression on the total of 78 samples of the Gamonedo cheese, cf. Ramos-Guajardo et al. (2019) Each entry refers to a different sample.

Expert 1	Expert 2
(65, 75, 85, 85)	(50, 50, 63, 75)
(35, 37, 44, 50)	(39, 47, 52, 60)
(66, 70, 75, 80)	(60, 70, 85, 90)
(70, 74, 80, 84)	(50, 56, 64, 74)
(65, 70, 75, 80)	(39, 45, 53, 57)
(45, 50, 57, 65)	(55, 60, 70, 76)
(60, 66, 70, 75)	(50, 50, 57, 67)
(65, 65, 70, 76)	(65, 67, 80, 87)
(60, 65, 75, 80)	(50, 50, 65, 75)
(55, 60, 66, 70)	(50, 55, 64, 70)
(60, 65, 70, 74)	(39, 46, 53, 56)
(30, 46, 44, 54)	(19, 29, 41, 50)
(60, 65, 75, 75)	(40, 47, 52, 56)
(70, 75, 85, 85)	(54, 55, 65, 76)
(44, 45, 50, 56)	(59, 65, 75, 85)
(51, 56, 64, 70)	(50, 52, 57, 60)
(40, 46, 54, 60)	(60, 60, 70, 80)
(55, 60, 65, 70)	(50, 54, 61, 67)
(80, 85, 90, 94)	(40, 46, 50, 50)
(80, 84, 90, 90)	(44, 50, 56, 66)
(65, 70, 76, 80)	(60, 64, 75, 85)
(75, 80, 86, 90)	(54, 56, 64, 75)
(65, 70, 73, 80)	(50, 50, 60, 66)
(70, 80, 84, 84)	(44, 46, 55, 57)
(55, 64, 70, 70)	(59, 63, 74, 80)
(64, 73, 80, 84)	(49, 50, 54, 58)
(50, 56, 64, 70)	(55, 60, 70, 75)
(55, 55, 60, 70)	(44, 47, 53, 60)
(60, 70, 75, 80)	(19, 20, 30, 41)
(64, 71, 80, 80)	(40, 44, 50, 60)
(50, 50, 55, 65)	(50, 50, 59, 66)
(50, 54, 60, 65)	(50, 53, 60, 66)
(65, 75, 80, 86)	(50, 52, 58, 61)
(50, 55, 60, 66)	(60, 65, 72, 80)
(40, 44, 50, 50)	(50, 50, 55, 60)
(70, 76, 85, 85)	(30, 34, 43, 47)
(44, 50, 53, 60)	(19, 25, 36, 46)
(34, 40, 46, 46)	(53, 63, 74, 80)
(40, 45, 51, 60)	
(84, 90, 95, 95)	

Simple calculations on data from Table 1 lead to means

$$\overline{x} = (57.65, 63.20, 69.18, 73.48)_T \quad \text{and} \quad \overline{y} = (47.34, 51.21, 59.87, 66.84)_T.$$

Substituting these results into (4) we obtain a value of our test statistic $t_0 = D(\overline{x}, \overline{y}) = 7.96$. Then, after combining samples and generating $K = 1000$ random permutations and following (6) we obtain the approximation of p-value of 0.002. Its interpretation is shown in Fig. 7, where one can find the histogram of all sampled differences $D(\overline{x^*}, \overline{y^*})$. Black dot indicates the value t_0 of the test statistic. The barely seen grey area on the right side of this dot corresponds to the probability of obtaining the distance between $\overline{x^*}$ and $\overline{y^*}$ not smaller than t_0. Therefore, we can rather confidently reject the null hypothesis and conclude that there is no general agreement between experts' opinions on the overall impression of the Gamonedo cheese.

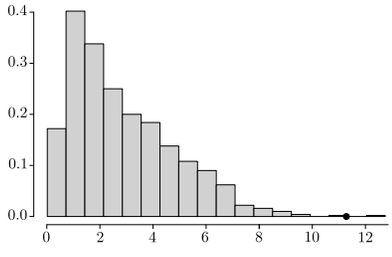


Fig. 7

The permutation agreement test considered above for two samples containing imprecise information is just one example of how fuzzy modeling can be combined with statistical inference. Although initially, some statisticians were skeptical about attempts to combine both theories, researchers realized that both statistics and fuzzy set theory should not be regarded as competitive, but that they can complement each other effectively. Moreover, expanding statistics with fuzzy sets not only solves some issues but also raises new questions. In particular, the distinction between the so-called ontic and epistemic sets yields different definitions of concepts as basic as variance and, consequently, different inferential tools. It is also worth noting that statisticians have also recognized fuzzy sets as convenient means for constructing procedures that allow the weakening of hypotheses or requirements that are excessively rigid.