CONSENSUS METHOD IN DATA PROCESSING

Yefim Bakman

Tel-Aviv University, *E-mail:* bakman@post.tau.ac.il

Abstract. The basic statistical method of data representation has not changed since its emergence in XIX century. Its simplicity was dictated by computational difficulties in the before-computers epoch. It turns out that such approach is not uniquely possible in the presence of quick computers. The suggested here method significantly improves data processing and graphical representation. It was also implemented in a computer program Consensus5 and verified through varied examples.

1. Introduction

The reader may ask why we need a new method if the old one is good – it builds histograms, calculates the average and the error. I offer to glance at Fig.1 where on the left a conventional histogram represents 250 normally distributed random numbers, whereas on the right the same data are presented by the computer program Consensus5. The difference is dramatic: the right graph is an ideal approximation to the gauss distribution, whereas the histogram is far from the normal curve. Such obvious advantage of the new method stems from the natural data representation.

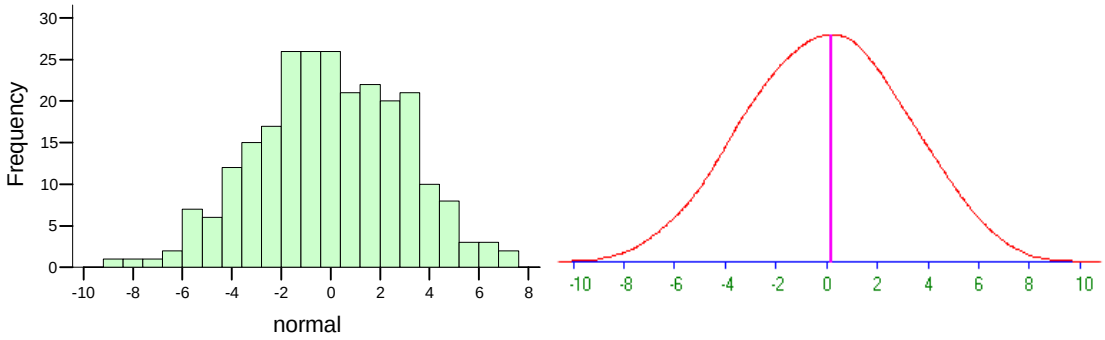

Fig.1. The graphical presentation of 250 normally distributed random numbers built by the computer program Consensus5 (on the right), and the conventional histogram (on the left).

What is even more important is that the consensus lets to discover erroneous data and faulty measuring devices as well. This is impossible in the standard approach context in which the instrumental error is not taken into account even though the term “measurement error” is used. Actually, it stands for the variability of the measurand, and these two different

concepts are blended in one by the standard approach. Later on we will discuss in more detail the distinction between different methods of data formation and, accordingly, different methods of their treatment. The latter will be illustrated by examples from varied fields.

2. The consensus method

It is frequently said that statistics is an exact science about non-precise things. Indeed, there exists discord between reality and the statistical method of its description. In the standard statistical approach an error may be infinitely large which evidences that the instrumental error is not implied. In reality each measuring device has its specification containing a commitment that the measured value x will not differ from the true value a by more than an admissible error δ: $|x-a|\leq\delta$. Thus, a measured value should be presented not by one number, but rather as minimum by the interval $[x-\delta, x+\delta]$, within which the true value a is located. If a device does not satisfy this requirement, it is considered *faulty*.

The main idea of the consensus method consists in that correct devices provide values x_i close to a, and consequently, to each other. More specific, correct intervals $[x_i-\delta, x_i+\delta]$ overlap because they share at least one common point a.

The common points which are shared by the greatest number of intervals are called here *consensus*. Consensus may contain one point or interval. In everyday life the word consensus means solidarity of opinions, but here we generalize this notion for the case of continuous numerical values presented in the form of intervals.

According to the definition, consensus includes the true value a (here we assume that the measurand variability may be neglected). And what about erroneous values? Their intervals do not intersect with the consensus which provides a means for their detection.

3. Improvement of the method for data representation

The representation of measured values in the form of intervals has the following imperfection – a tiny change in the measured value may bring to modification of the result due to the sharp break of the interval. The interval must not end abruptly. Such method for representation of non-precise data has been already developed in fuzzy numbers theory and is called *membership function* (see [1, 2]). Fig.2 demonstrates an example of such function having a trapezoidal shape. The value $A(x)$ of this function is the confidence that the measured quantity is equal to x. For example, from Fig.2 one concludes that $A(x-\delta)=1$ just like $A(x)=1$ or $A(x+\delta)=1$, i.e. a deviation from the measured value within the error limits

does not alter the confidence. However A(x+1.5δ)=0.5 and then the confidence gradually falls till zero at x+2δ remaining continuous all the time.

From the viewpoint of such representation the standard method presents a degenerate case in which the trapezoid transforms into vertical segment, i.e. the measurement error is equal to zero.

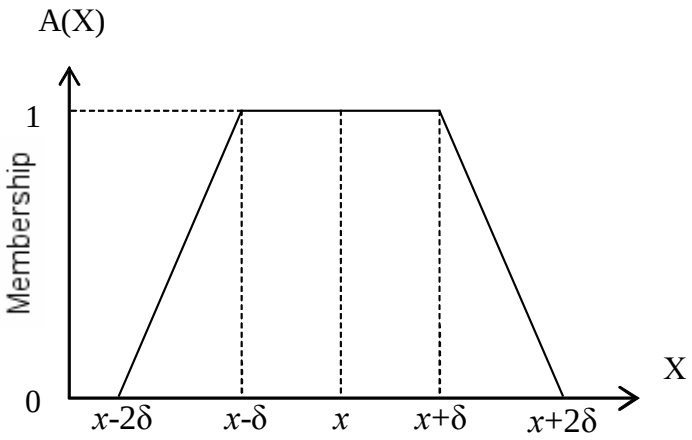

Fig.2. An example of the membership function.

When n measurements of the same quantity are available, the composite membership function is the sum of the separate functions divided by n: $A(x) = \sum_{i=1}^{n} A_i(x)/n$. The advantage of such representation can be seen if applied to an example. Suppose that a quantity was measured by six devices having δ=0.2, and the following values were obtained 0.9, 1, 1.1, 3, 4, 5. The composite membership function for these six values is shown in Fig. 3.

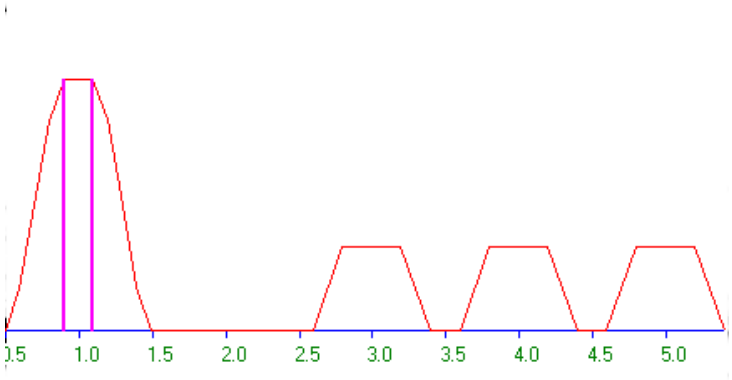

Fig. 3. The composite membership function for the six values.

The first three values overlap forming a high trapezoid. Because of the improvement of non-precise data representation it is also necessary to specify the notion of consensus:

Consensus is a point or an interval, in which the composite membership function has the largest value.

Following this definition, we conclude that for our case the consensus constitutes the interval [0.9, 1.1] marked by the vertical lines in Fig.3. Actually, a membership function graph can serve as an improved substitution for histogram which was decisively shown by Fig.1.

4. Consensus and erroneous data detection

The idea of consensus can be better expressed as follows: accurate values are close to the true value, consequently, they are close to each other and their intervals overlap forming consensus. On the contrary, erroneous values are scattered around randomly. They are far from each other having low chances to overlap; therefore they cannot change the consensus. This means that consensus is robust against adding erroneous data; it also helps to reveal erroneous values. Fig.3 shows that the values 3, 4, 5 are out of the consensus; consequently, they are erroneous and should be expelled from the consideration.

The situation with faulty measuring devices is even better. Suppose the chance for a casual value to overlap with the consensus is equal to p. Then the probability for a faulty measurer to overlap twice with the consensus is equal to p^2. If we take p=0.1, then p^2 = 0.01 is practically negligible. Therefore the more different quantities are measured by the same measurers, the more chances to detect defective devices. This rule is demonstrated in section 5 in which unskilled experts are reliably detected by means of a questionnaire containing nine questions.

The idea may be illustrated graphically if only two quantities X and Y were measured by six measuring devices S1-S6, the measurement error being specified as 0.2. The obtained values are presented in Table 1.

	S1	S2	S3	S4	S5	S6
X	1.9	2	2.1	4	6	7
Y	0.9	1	1.1	1	5	4

Table 1. Six values of X and Y as measured by the devices S1-S6.

In this example the faultiness of S4 cannot be detected by y_4 value only, because y_4 is occasionally inside the Y-consensus, however x_4 helps to do it. The two dimensional membership function for the six measured values S1-S6 is shown in Fig. 4. The highest truncated pyramid (it is also the nearest one) is formed by the imposition of the separate functions of the measurements S1-S3, while the three low pyramids correspond to the measurements S4-S6. Those latter do not overlap.

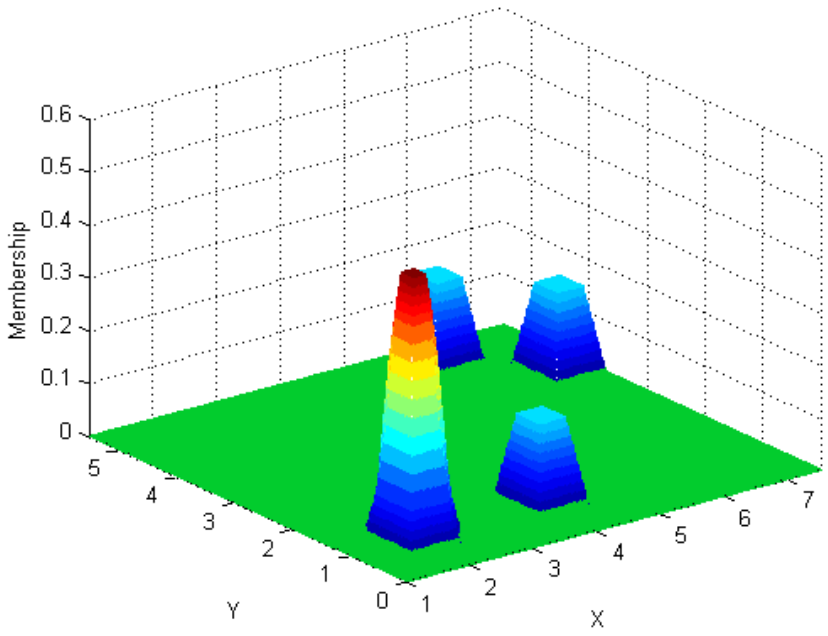

Fig.4. The composite membership function for the data from Table 1.

It is seen from Fig.4 that the measurers S1-S3 possess a common area in the vicinity of the point (2;1), where the membership function has its peak value. The square $1.9 \le x \le 2.1;\ 0.9 \le y \le 1.1$ constitutes the consensus, in other words, the coordinates of any points of this square can be the true values of X and Y. The remaining three pyramids (including S4) are out of the consensus; hence they represent defective measuring devices. Thus, measuring two quantities X and Y is preferable for detecting faulty devices than measuring only one of them.

It should be noted that the standard method does not provide a histogram analogy for two variables, whereas the two-dimensional membership function does the job.

5. Computer program Consensus5 testing

Among others the program was tested by two surveys. The respondents answered 9 questions (estimate on the scale 1 to 7 the climate of Israel, the public health system, the economic status and so on).

After the consensus was calculated for every question of the questionnaires, four virtual respondents with random estimates were added to the surveys. The program evaluated them as not competent, and the calculated consensus coincided with the first run value (see Fig.5). Thus the consensus proved its robustness.

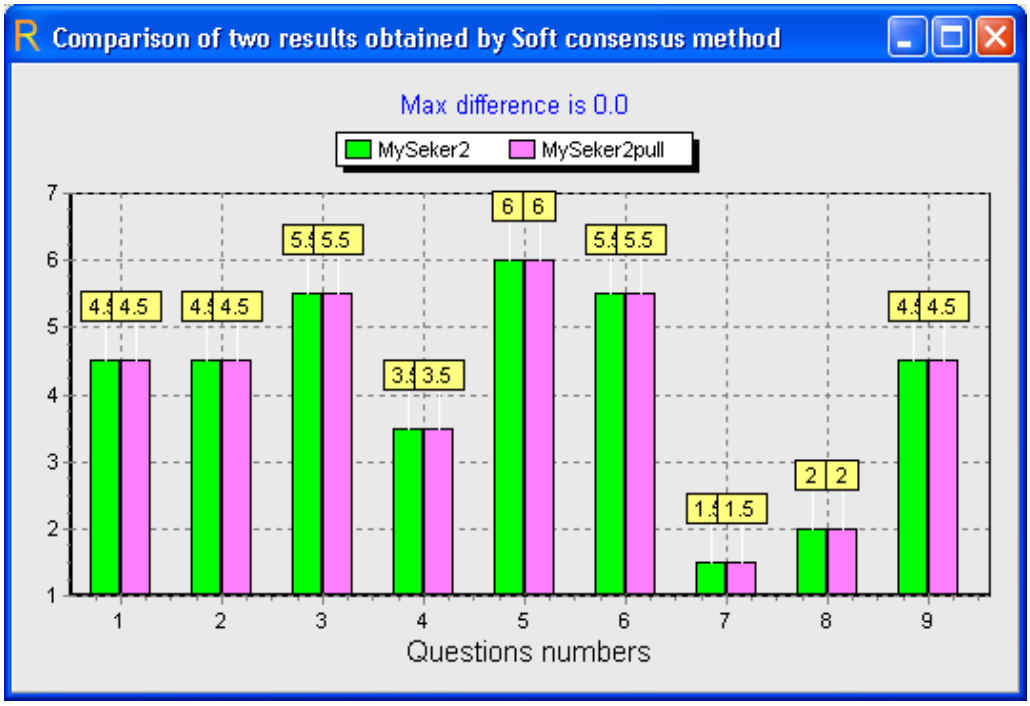

Fig.5. After addition of four artificial respondents with random estimates to the survey the program Consensus5 labeled them as erroneous and deleted them from the data. It is natural that the consensus remained unchanged.

The sample means and medians were also calculated for the surveys (see Fig.6). The greatest difference of the sample means between the two runs was 1.2 points, whereas for the medians the maximum difference was 3 points which evidences instability of these two parameters against adding erroneous data to the surveys.

It should be noted that no specifications were attached to experts in the previous example. Nevertheless a researcher is free to choose his/her own requirements to a competent expert from logical considerations, for instance, the error must not exceed 0.5. Indeed, since the estimations are integers, if an expert's choice is 3.5, he/she is forced to choose 3 or 4 that

is to deviate by 0.5 from the genuine value. Hence 0.5 is the minimal measurement error for such questionnaires.

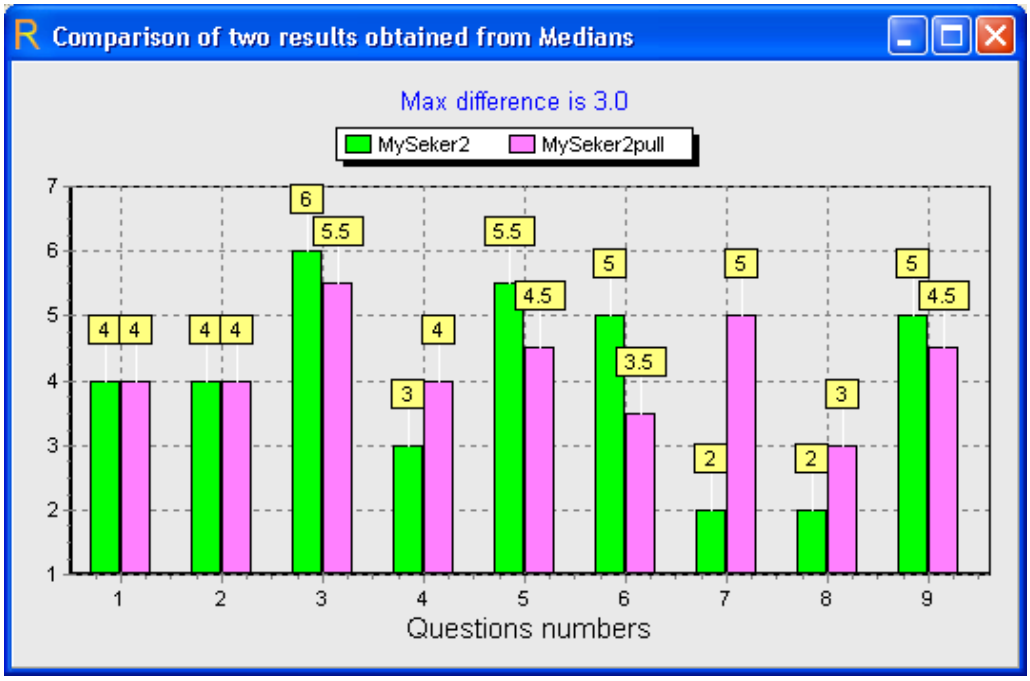

Fig.6. After addition of four artificial respondents with random estimates to the survey the greatest difference of the sample medians composed 3 points. This evidences that the median is not robust against adding erroneous data to the surveys.

Conclusion

Testing the computer program Consensus5 demonstrated the following advantages of the suggested method for data processing:

a) Felicitous graphical representation even for small samples.
b) Consensus is robust against erroneous measurements.
c) Consensus helps to detect erroneous value and faulty measuring devices.

Thus, all the bundle of problems associated with the standard statistical approach was solved with the help of one unique correction of the method for data representation. In the light of the suggested correction the old method appears to be a particular case, in which the interval of possible values degenerates into a point, herewith the consensus turns into the mode (the most frequent value). Such representation would be true if all measurements were absolutely accurate which is unreal. It is probable that the choice of the oversimplified method for data representation was dictated by the computational difficulties about two centuries ago. Nowadays we can afford the luxury of the choice of the pertinent algorithm for data processing not being restricted by the number of computations.

I invite everyone to send me his/her data for further testing the program Consensus5. Please write the word "Consensus" as the subject of the electronic message.

References

[1] G. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic, Theory and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[2] J. Buckley. *Fuzzy Statistics*, Springer, NY, 2004.