6-1-2022

How to Combine Expert Estimates? How to Estimate Probability in the Intersection of Two Populations?

Miroslav Svitek
Czech Technical University in Prague, svitek@fd.cvut.cz

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:

Technical Report: UTEP-CS-22-77

How to Combine Expert Estimates? How to Estimate Probability in the Intersection of Two Populations?

Miroslav Svítek, Olga Kosheleva, Vladik Kreinovich, and Nguyen Hoang Phuong

Abstract In this paper, we consider two different practical problems that turned to be mathematically similar: (1) how to combine two expert-provided probabilities of some event and (2) how to estimate the frequency of a certain phenomenon (e.g., illness) in an intersection of two populations if we know the frequencies in each of these populations. In both cases, we use the maximum entropy approach to come up with a solution.

1 Formulation of the First Problem

The problem. Suppose that the experts E_1 and E_2 provided two estimates p_1 and p_2 for the probability of some event E. We would like to provide a single estimate that takes both estimates into account.

What information we can use. To properly combine the two estimates, it is important to take into account how related are the opinions of the two experts.

- If in all previous situations and in this situation, the experts gave almost identical opinions, this probably means that they use the same technique to provide their estimates. In this case, the opinion of the second expert does not add anything

Miroslav Svítek
Faculty of Transportation Sciences, Czech Technical University in Prague, Konviktska 20
CZ-110 00 Prague 1, Czech Republic, e-mail: svitek@fd.cvut.cz

Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso, 500 W. University, El Paso, Texas 79968, USA
e-mail: olgak@utep.edu, vladik@utep.edu

Nguyen Hoang Phuong
Division Informatics, Math-Informatics Faculty, Thang Long University, Nghiem Xuan Yem Road
Hoang Mai District, Hanoi, Vietnam, e-mail: nhphuong2008@gmail.com

new to the opinion of the first expert. So, the combined probability will still be the same value p_1.
- If in the previous situations, the experts' opinions were independent, this means that they use different data and different techniques to estimate the probability. In this case, e.g., if both experts believe that this event is possible, then taking both opinions into account should increase this probability.
- On the other hand, if the expert opinions are negatively correlated, then we do not know whom to believe, and we should not take either of the probabilities seriously. In this case, the combined probability should be close to the do-not-know 0.5 value.

What we do in this paper. In this paper, we show how to come up with a reasonable numerical value of the combined probability.

2 Formulation of the Second Problem

Formulation of the problem. Suppose that we know the frequencies of a certain phenomenon in two different populations: e.g., the frequency of a certain disease in a 50-60 age group and the frequency of this disease in women. Based on these two frequencies, what is the reasonable estimate for the frequency of this disease in the intersection of these two populations – i.e., among women in the 50-60 age bracket; see, e.g., [1] and references therein.

What we do in this paper. In this paper, we show that this problem is mathematically similar to the first one, and thus, all the methods for solving the first problem can be automatically applied to the second problem as well.

3 Formulation of the Problems in Precise Terms

Notations: case of the first problem. In the first problem, let us make the following notations for the random variables:

- Let E be equal to 1 if the event happens and 0 if it does not.
- Let e_1 be equal to 1 if the first expert is correct in a randomly selected situation, and 0 if the first expert is wrong.
- Similarly, let e_2 be equal to 1 if the second expert is correct in a randomly selected situation, and 0 if the second expert is wrong.

What we know and what we want to estimate. What we know:

- We know the conditional probabilities $p(E \mid e_1) = p_1$ and $p(E \mid e_2) = p_2$.
- Based on the analysis of the previous expert opinions, we can estimate the probability $p(e_1)$ – by counting how many times the first expert was right.

- Similarly, based on the analysis of the previous expert opinions, we can estimate the probability $p(e_2)$ – by counting how many times the second expert was right.
- We can also estimate the probability $p(e_1 \,\&\, e_2)$ that both experts were right.

Based on this information, we want to estimate the probability $p(E)$.

What if we have only one expert? In this case, we know the conditional probability $p_1 = p(E \mid e_1)$ and the probability $p(e_1)$, and we want to estimate the probability $p(E)$.

Second problem. In the second problem, we only consider folks who belong to one (or both) of the two populations. In this case, we consider the following random variables:

- Let E be equal to 1 if a randomly selected element has the desired phenomenon.
- Let e_1 be 1 if a randomly selected element belongs to the first population.
- Let e_2 be 1 if a randomly selected element belongs to the first population.

For this problem, we know the following information:

- We know the frequencies $p(E \mid e_1) = p_1$ and $p(E \mid e_2) = p_2$ of the phenomenon in each population.
- We know the number of elements n_1 in the first population, the number of elements n_2 in the second population, and the number of elements n_{12} that belong to both populations. In this case, the overall number of elements that belongs to both populations is equal to $n_1 + n_2 - n_{12}$.
- Thus, we can estimate the probabilities of e_1, e_2, and $e_1 \,\&\, e_2$ as follows:

$$p(e_1) = \frac{n_1}{n_1 + n_2 - n_{12}}; \quad p(e_2) = \frac{n_2}{n_1 + n_2 - n_{12}}; \quad p(e_1 \,\&\, e_2) = \frac{n_{12}}{n_1 + n_2 - n_{12}}.$$

Based on this information, we want to find the probability $p(E \mid e_1 \,\&\, e_2)$.

4 How These Problems Can Be Solved

Let us use the Maximum Entropy approach. Situations in which we only have partial information about probabilities – and thus, several different probability distributions are consistent with our knowledge – are ubiquitous. In such cases, it makes sense not to pretend that our uncertainty is low – and thus, to select the distribution with the largest possible uncertainty.

A natural measure of uncertainty of a probability distribution is the average number of binary ("yes"-"no") questions that we need to ask to fully determine which statements are true and which are false. It is known that this number is equal to Shannon's entropy $S = -\sum P_i \cdot \log_2(P_i)$, where P_i are the probabilities of different possible situations; see, e.g., [2]. Thus, we need to select the distribution with the largest possible entropy. Such a selection is known as the Maximum Entropy approach.

What this means for the first problem. In the first problem, we have three basic statement E, e_1, and e_2. Each of these statements is either true or false. Thus, we have $2^3 = 8$ possible situations:

$$E \,\&\, e_1 \,\&e_2,\;\; E \,\&\, e_1 \,\&\neg e_2,\;\; E \,\&\, \neg e_1 \,\&e_2,\;\; E \,\&\, \neg e_1 \,\&\neg e_2,$$

$$\neg E \,\&\, e_1 \,\&e_2,\;\; \neg E \,\&\, e_1 \,\&\neg e_2,\;\; \neg E \,\&\, \neg e_1 \,\&e_2,\;\; \neg E \,\&\, \neg e_1 \,\&\neg e_2.$$

Let us use the following notations for their probabilities:

$$p_{111} \stackrel{\text{def}}{=} E \,\&\, e_1 \,\&e_2,\;\; p_{110} \stackrel{\text{def}}{=} E \,\&\, e_1 \,\&\neg e_2,\;\;\; p_{101} \stackrel{\text{def}}{=} E \,\&\, \neg e_1 \,\&e_2,$$

$$p_{100} \stackrel{\text{def}}{=} E \,\&\, \neg e_1 \,\&\neg e_2,\;\; p_{011} \stackrel{\text{def}}{=} \neg E \,\&\, e_1 \,\&e_2,\;\;\; p_{010} \stackrel{\text{def}}{=} \neg E \,\&\, e_1 \,\&\neg e_2,$$

$$p_{001} \stackrel{\text{def}}{=} \neg E \,\&\, \neg e_1 \,\&e_2,\;\; p_{000} \stackrel{\text{def}}{=} \neg E \,\&\, \neg e_1 \,\&\neg e_2.$$

These eight probabilities must add up to 1:

$$p_{111} + p_{110} + p_{101} + p_{100} + p_{011} + p_{010} + p_{001} + p_{000} = 1. \tag{1}$$

Since we know the probability $p(e_1)$, the fact that we know the value $p_1 = p(E \mid e_1) = p(E \,\&\, e_1)/p(e_1)$ is equivalent to knowing the probability $p(E \,\&\, e_1) = p_1 \cdot p(e_1)$. In terms of the basic probabilities, the probability $p(E \,\&\, e_1)$ has the form

$$p(E \,\&\, e_1) = p(E \,\&\, e_1 \,\&\, e_2) + p(E \,\&\, e_1 \,\&\, \neg e_2) = p_{111} + p_{110}.$$

Thus, we have

$$p_{111} + p_{110} = p_1 \cdot p(e_1). \tag{2}$$

Similarly, since we know the probability $p(e_2)$, the fact that we know the value $p_2 = p(E \mid e_2) = p(E \,\&\, e_2)/p(e_2)$ is equivalent to knowing the probability $p(E \,\&\, e_2) = p_2 \cdot p(e_2)$. In terms of the basic probabilities, the probability $p(E \,\&\, e_2)$ has the form

$$p(E \,\&\, e_2) = p(E \,\&\, e_1 \,\&\, e_2) + p(E \,\&\, \neg e_1 \,\&\, e_2) = p_{111} + p_{101}.$$

Thus, we have

$$p_{111} + p_{101} = p_2 \cdot p(e_2). \tag{3}$$

Information about the values $p(e_1)$, $p(e_2)$, and $p(e_1 \,\&\, e_2)$ takes the following form:

$$p_{111} + p_{110} + p_{011} + p_{010} = p(e_1); \tag{4}$$

$$p_{111} + p_{101} + p_{011} + p_{001} = p(e_2); \tag{5}$$

$$p_{111} + p_{011} = p(e_1 \,\&\, e_2). \tag{6}$$

So, to find the values p_{ikj}, we need to maximize the entropy

$$S = -p_{111} \cdot \log_2(p_{111}) - p_{110} \cdot \log_2(p_{110}) - p_{101} \cdot \log_2(p_{101}) -$$

$$p_{100} \cdot \log_2(p_{100}) - p_{011} \cdot \log_2(p_{011}) - p_{010} \cdot \log_2(p_{010}) -$$
$$p_{001} \cdot \log_2(p_{001}) - p_{000} \cdot \log_2(p_{000}) \quad (7)$$

under the constraints (1)–(6).

Entropy (6) is a convex function of the probabilities, and the constraints are linear in terms of these probabilities. Thus, we can use the feasible convex optimization algorithms to find the desired probabilities; see, e.g., [2, 3]. Once we find all the probabilities p_{ijk}, we can compute the desired probability $p(E)$ as

$$p(E) = p_{111} + p_{110} + p_{101} + p_{100}. \quad (8)$$

Comments. We can similarly consider the case when we have more than two experts and the case when we only have one expert. In the situation when we have only one expert, we have four possible situations

$$E \,\&\, e_1, \;\; E \,\&\, \neg e_1, \;\; \neg E \,\&\, e_1, \;\; \neg E \,\&\, \neg e_1.$$

Let us denote the probabilities of these situations by

$$p_{11} = p(E \,\&\, e_1), \;\; p_{10} = p(E \,\&\, \neg e_1), \;\; p_{01} = p(\neg E \,\&\, e_1), \;\; p_{00} = p(\neg E \,\&\, \neg e_1).$$

These four probabilities must add up to 1:

$$p_{11} + p_{10} + p_{01} + p_{00} = 1. \quad (9)$$

The available information – i.e., the values $p(e_1)$ and p_1 – lead to the following constraints:

$$p_{11} + p_{01} = p(e_1) \quad (10)$$

and

$$p_{11} = p_1 \cdot p(e_1). \quad (11)$$

In this case, we know the values p_{11} – from the equation (11) – and we can, thus, determine the probability p_{01} from the formula (10), as

$$p_{01} = p(e_1) - p_1 \cdot p(e_1) = p(e_1) \cdot (1 - p_1).$$

The only constraint on the remaining two values p_{00} and p_{10} (coming from the condition (9)) is that

$$p_{00} + p_{10} = 1 - (p_{01} + p_{11}) = 1 - p(e_1).$$

In this case, the maximum entropy approach leads to equal values of these two probabilities:

$$p_{00} = p_{10} = \frac{1 - p(e_1)}{2}.$$

Thus, the resulting estimate for the desired probability $p(E) = p_{11} + p_{10}$ has the form

$$p(E) = p_1 \cdot p(e_1) + \frac{1 - p(e_1)}{2} = \frac{1}{2} + p(e_1) \cdot \left(p_1 - \frac{1}{2}\right). \quad (12)$$

This formula can be alternatively reformulated as

$$p(E) - \frac{1}{2} = p(e_1) \cdot \left(p_1 - \frac{1}{2}\right). \quad (13)$$

In other words, we should not take the expert estimate p_1 at face value, we should adjust this estimate based on the expert's track record.

What this means for the second problem. From the mathematical viewpoint, the two problems has similar inputs, the only two differences are as follows:

- first, we only consider objects that belong to one of the populations, so
$$p_{000} = p_{100} = 0; \quad (14)$$
- second, what we want to estimate is different: instead of the probability $p(E)$ (as in the the first problem) we want to estimate the conditional probability $p(E \mid e_1 \,\&\, e_2)$.

Thus, to solve the second problem, we perform the same optimization as in the first problem – with the additional constraint (14) – to find the probabilities p_{ijk}, and then estimate:

$$p(E \mid e_1 \,\&\, e_2) = \frac{p(E \,\&\, e_1 \,\&\, e_2)}{p(e_1 \,\&\, e_2)} = \frac{p_{111}}{p_{011} + p_{111}}. \quad (15)$$

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

References

1. P. Baldi, "What's hot in uncertain reasoning", *The Reasoner*, 2022, Vol. 16, No. 3, pp. 23–24.

2. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
3. R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1997.