How to Estimate Unknown Unknowns: From Cosmic Light to Election Polls

Talha Azfar[1], Vignesh Ponraj[2], and Vladik Kreinovich[2]
Departments of [1]Electrical and Computer Engineering and [2]Computer Science
University of Texas at El Paso, El Paso, Texas 79968, USA
tazfar@miners.utep.edu, vponraj@miners.utep.edu,
vladik@utep.edu

First case study: space light. Stars in the galaxies emit light. Some galaxies are visible, others are not too far away to be visible individually, but since there are many of them, together they contribute to the optical background visible by space telescopes. We have a reasonably good understanding of how galaxies are distributed in space and what light an average galaxy emits. Based on this information, we can estimate the amount of background light. Interestingly, the observed amount is almost exactly twice larger than the estimate – meaning that there are some additional sources of light in the Universe; see, e.g., [1].

Second case study: election polls. It is known, from statistics, that if we estimate the probability of an event based on the sample of size n, then the standard deviation σ of the corresponding accuracy is equal to $\sqrt{p \cdot (1-p)/n}$. In particular, when we use the poll of $n = 1000$ randomly selected people to estimate the probability p of a candidate's win, then for candidates with approximately equal chances, where $p \approx 0.5$, we get $\sigma \approx 1.7\%$. So, with 95% confidence, this should estimate the probability with $2\sigma \approx 3.5\%$ accuracy. In practice, the largest deviation is twice larger; see, e.g., [2].

How can we explain these two facts. In both cases, taking unknown unknowns into account doubles the corresponding value. How can we explain that?

In both cases, we know the estimated value v, and we want to estimate the actual value a. The only information that we have about a is that $a > v$. Based on this information, how can we estimate a? To answer this question, let us consider the unknown value a as the new measuring unit for the corresponding quantity. In terms of this new unit, the value a will take the form $A = 1$, and the value v will have the form $V = v/a$. Since $0 < v < a$, we have $0 < V < 1$, so the only thing we know about this value V is that it is located on the interval $[0, 1]$. We have no reason to assume that some of these values are more probable than others, so it makes sense to assume that all these values are equally probable – this argument is known as Laplace Indeterminacy Principle. In other words, it is reasonable to assume that the value V is uniformly distributed on the interval $[0, 1]$.

If we want to represent this distribution by a single number, a reasonable choice is to select the value V_s for which the mean square deviation from the actual (unknown) value v is the smallest possible. One can easily check that this V_s is the mean value of V, i.e., $V_s = 1/2$. Thus, we have $v/a = 1/2$. Based on this relation, if we know v, then a reasonable estimate for a is $a = 2v$ – which is exactly what we observe in the above two cases.

References

[1] J. L. Bernal, G. Sato-Polito, and M. Kamionkowski, "Cosmic optical background excess, dark matter, and line-intensity mapping", *Physical Review Letters*, 2022, Vol. 129, Paper 231301.

[2] H. Shirani-Mehr, D. Rotschild, S. Goel, and A. Gelman, "Disentagling bias and variance in electrion polls", *Journal of the American Statistical Association*, 2018, Vol. 113, No. 522, pp. 607–614.