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

Talha Azfar[1], Vignesh Ponraj[2], Vladik Kreinovich[2], and Nguyen Hoang Phuong[3]

Departments of [1]Electrical and Computer Engineering and [2]Computer Science
University of Texas at El Paso, 500 W. University, El Paso, Texas 79968, USA
tazfar@miners.utep.edu, vponraj@miners.utep.edu, vladik@utep.edu

[3]Artificial Intelligence Division, Information Technology Faculty
Thang Long University, Nghiem Xuan Yem Road
Hoang Mai District, Hanoi, Vietnam, nhphuong2008@gmail.com

1. General introduction

- In two different areas of study – the study of space light and the study of elections – there is a similar puzzling phenomenon.
- The observed value of the corresponding quantity is exactly twice larger that reasonable models predict.
- In this talk, we provide a possible common explanation for these two phenomena.

2. What is space light

- Many celestial objects emit light.
- This is why we see many stars and galaxies, this is why other stars and galaxies can be seen via telescopes.
- What we see is the light they emit.
- Usually, astronomers study light from visible stars and galaxies, where we can see the corresponding object.
- Some galaxies are too far away to be visible individually.
- However, since there are many of them, they contribute to the optical background that is visible by space telescopes.
- This background is known as *space light*.

3. We can estimate the expected amount of space light

- We have a reasonably good understanding of:
  - how galaxies are distributed in space and
  - what amount of light an average galaxy emits.
- Based on this information, we can estimate the amount of background light.

4. The observed amount of space light is twice larger than expected

- Interestingly, the observed amount is almost exactly twice larger than the estimate.
- This means that there are some additional sources of light in the Universe.
- That there are some unexpected sources of slight is natural.
- However, the fact that the observed amount of light is exactly twice larger than expected deserves explanation.
- In this talk, we provide:
  - a natural explanation for this empirical fact
  - as well as for the similar empirical fact about elections.

5. Election polls: reminder

- To get a good understanding of how people will vote, specialists ask a random sample of people how they will vote in the forthcoming elections.
- This process is known as *election polls*.
- After the poll:
  - the percentage of people who expect to vote for a certain candidate
  - is used as a reasonable approximation for the percentage of people who will actually vote for this candidate.
- Of course, percentages based on a small sample are only an approximation to the overall percentages.
- A natural question is: how accurate are the polls?
- If, based on a poll, one candidate is several points ahead, how confident are we that this candidate will win?

6. How is the accuracy of election polls usually estimated?

- 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.

7. Observed standard deviation is exactly twice larger

- In practice, the largest deviation is twice larger then what we would expect.
- That standard deviation is larger than expected is natural.
- People change their opinions, and this adds to the difference between how people answer in the poll and how they actually vote.
- However, the fact that the observed standard deviation is exactly twice larger than expected deserves explanation.
- In this talk, we provide:
  - a natural explanation for this empirical fact,
  - as well as for the similar empirical fact about space light.

8. Analysis of the problem

- In both case studies, taking unknown unknowns into account doubles the corresponding value.
- How can we explain that?
- In both case studies:
  - 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 $a \geq v$.
- Based on this information, how can we estimate a?

9. Let us reformulate the problem, to make it easier to answer: idea

- In the above formulation, we have two real numbers: v and a.
- To simplify the problem, let us take into account that the numerical value of each quantity depends on the selection of a measuring unit.
- For example, the same height of 1.7 meters takes the value 170 if we use centimeter as a measuring unit.
- Let us use this idea to simplify our problem.
- For this purpose, let us select 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 $V = v/a$.

10. Let us reformulate the problem (cont-d)

- Thus, the above problem is reformulated as follows:
  - we know that in the new unit, the actual value is 1, and
  - we want to find the value V that described the estimated value in terms of this new unit.
- The only information that we have about the desired value V is:
  - that $0 < V \leq 1$, i.e.,
  - that the value V is that it is located on the interval $[0, 1]$.

11. It is natural to use Laplace Indeterminacy Principle

- 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.
- Based on this argument, we conclude that the value V is uniformly distributed on the interval $[0, 1]$.

12. From distribution to a single numerical estimate

- We have a reasonable distribution of the set of all possible values V.
- What we want, however, is a single numerical estimate.
- In general, 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., $V_s = 0.5$.

13. This conclusion indeed explains the above phenomena

- We have $v/a = 1/2$.
- Based on this relation:
  - if we know v,
  - then a reasonable estimate for a is $a = 2v$.
- This is exactly what we observe in the above two case studies.

14. Acknowledgments

This work was supported in part by:

- National Science Foundation grants 1623190, HRD-1834620, HRD-2034030, and EAR-2225395;
- AT&T Fellowship in Information Technology;
- program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- a grant from the Hungarian National Research, Development and Innovation Office (NRDI).