For statistical analysis of big data, interval uncertainty is needed

Olga Kosheleva[1] and Vladik Kreinovich[2]

Departments of [1]Teacher Education and [2]Computer Science
University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Statistical testing: a brief reminder

- A usual statistical approach to processing data $x_1, \ldots, x_n$ is as follows:
  - we come up with a model – e.g., based on the training part of the data – and
  - then we test whether this model adequately describes the remaining testing part of the data.
- For example, it may turns out that the observations are consistent with a normal distribution with mean m and standard deviation σ.
- Then we can use the chi-square criterion and check whether, or appropriate values $\chi^2_{n,1-\alpha}$ and $\chi^2_{n,\alpha}$, we have

$$\chi^2_{n,1-\alpha} \le \sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^2} \le \chi^2_{n,\alpha}.$$

2. Statistical testing: a brief reminder (cont-d)

- These tests are designed in such a way that:
  - when the actual distributions is the assumed one,
  - this test returns "true" with frequency close to 1.
- Also, when the actual distribution is different from the assumed one, then:
  - for sufficiently large n – above a certain threshold n_0 –
  - the corresponding test fails with frequency close to 1.
- This traditional statistical approach has worked successfully for more than a century.

3. Formulation of the problem

- However:
  - with the emergence of big data, when we have millions and even billions of data points,
  - this traditional statistical approach often fails.
- The reason for this failure is clear.
- In most applications areas – e.g., in econometrics – all statistical models are approximate.
- When n was reasonably small, much smaller than the threshold value n_0, the tests still worked.
- However, big data often means $n > n_0$, so the tests fail.
- As a result:
  - either we cannot find a model that fits the training data,
  - or, if such a model is found, we cannot show that it fits the testing data.

4. Formulation of the problem (cont-d)

- This phenomenon is known as *macronumerosity*.
- This is well-known problem in many application areas – e.g., in modeling climate change.

5. Statistical approach to this problem does not work

- That the model is approximate means that there are some close values $\widetilde{x}_i \approx x_i$ that fit this model.
- A typical statistical idea would be to find the distribution for the approximation error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.
- However, in this case, we would still assume some exact distribution for x_i.
- So this brings us back to the same problem.

6. Intervals form a natural solution

- From the interval viewpoint, a natural solution out of this seemingly vicious circle is:
  - not to assume any specific distribution for Δx_i, but
  - instead to use interval uncertainty, i.e., to assume that all the values Δx_i are within an interval $[-\Delta, \Delta]$.
- In this case:
  - for each model and each corresponding test $C(x_1, \ldots, x_n) \le C_0$ that is not satisfied for the actual data,
  - we can describe the degree to which data fits the model.
- We can do it by computing the smallest Δ for which some Δ-close values $\widetilde{x}_i$ satisfy the test.
- Usually, the values Δx_i are small, so we can ignore terms quadratic in Δx_i.

7. Intervals form a natural solution (cont-d)

- In this linear approximation, we have

$$\widetilde{C} \stackrel{\text{def}}{=} C(\widetilde{x}_1, \ldots, \widetilde{x}_n) =$$

$$C(x_1 + \Delta x_1, \ldots, x_n + \Delta x_n) = C(x_1, \ldots, x_n) + \sum_{i=1}^{n} C_{,i} \cdot \Delta x_i.$$

- Here $C_{,i}$ are partial derivatives of C with respect to x_i:

$$C_{,i} \stackrel{\text{def}}{=} \frac{\partial C}{\partial x_i}.$$

- We want to make sure that $\widetilde{C} \le C_0$, i.e., that

$$\sum_{i=1}^{n} C_{,i} \cdot (-\Delta x_i) \ge C(x_1, \ldots, x_n) - C_0.$$

8. Intervals form a natural solution (cont-d)

- When $|\Delta x_i| \le \Delta$, the largest value of the sum $\sum\limits_{i=1}^{n} C_{,i} \cdot (-\Delta x_i)$ is

$$\sum_{i=1}^{n} |C_{,i}| \cdot \Delta = \Delta \cdot \sum_{i=1}^{n} |C_{,i}|.$$

- Thus, the smallest Δ for which we can get $\widetilde{C} \le C_0$ is equal to

$$\Delta = \frac{|C(x_1, \ldots, x_n) - C_0|}{\sum\limits_{i=1}^{n} |C_{,i}|}.$$

- Then, e.g., between two models with the same number of parameters, we can select the model with the smallest Δ.

9. References

- A. Nichols and M. E. Schaffer, "Practical steps to improve specification testing", In: N. N. Thach, D. T. Ha, N. D. Trung, and V. Kreinovich (eds.), *Prediction and Causality in Econometrics and Related Topics*, Springer, Cham, Switzerland, 2022, pp. 75–88.
- M. E. Schaffer, "Null hypothesis misspecification testing revisited: how (not) to test orthogonality conditions", In: V. Kreinovich, W. Yamaka, and S. Leurcharusmee (eds.), *Data Science for Econometrics and Related Topics*, Springer, Cham, Switzerland, to appear.

10. Acknowledgments

This work was supported in part:

- by the US National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
  - HRD-1834620 and HRD-2034030 (CAHSI Includes),
  - EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES),
- by the AT&T Fellowship in Information Technology, and
- by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,