Statistics for Imprecise Data: the Key for Enlarging the IP Community

Marco de Angelis and Scott Ferson MARCO.DE-ANGELIS,FERSON[AT]LIVERPOOL.AC.UK
University of Liverpool, Institute for Risk and Uncertainty, Liverpool, UK

Luke Green LUKE[AT]DATACLIMATE.CO
Vivaldi Analytics, Stony Brook, New York, USA

Poster Abstract

As a discipline, the theory of imprecise probabilities may be pricing itself out of the market in the sense that its complexity, computational burden, and requisite mathematical sophistication required for nontrivial applications are prohibitive in many subject domains. For the discipline to grow, it is essential to foster broad interest and use across science and engineering. This will involve recruiting a class of users who may not develop methods but who will apply them in their routine work. Their applications give evidence of the utility of the imprecise probabilities approach and its underlying philosophy. This implies that someone who sees imprecision in a data set but who lacks special training in uncertainty quantification or imprecise probabilities should be able to apply convenient algorithms for basic statistics.

When the data set has imprecision, computing statistics can be challenging. For example, for data in the form of intervals, using naïve interval analysis yields results with inflated uncertainty because of repetitions of variables in the formulas. Moreover, finding optimal bounds on many basic statistics are NP-hard problems that grow in difficulty with the size of the data set. It is practically impossible to solve these problems for large data sets with a simple sampling strategy, such as Monte Carlo, in which the formula for the variance is treated like a black box evaluated for many possible configurations of the data points within their respective intervals.

Over the last century, statistics has focused on developing methods for analyses in which data sample size is limiting. But not all uncertainty in data has to do with small sample sizes. Although most statistical analyses today ignore the uncertainty reported by laboratories and empiricists as interval measurement uncertainty, this is clearly not always because this uncertainty is negligibly small. We believe it may instead be due to the lack of friendly software to handle it. We announce and describe a software library that is intended to provide convenient access to basic statistics for interval and censored data. The library of algorithms is being used to develop on-line and stand-alone software for analyzing data sets containing imprecision as well as sampling uncertainty. The algorithms in the library require users to make fewer dubious assumptions about the data set than currently popular methods for handling data censoring, missingness, and lack of independence. The library currently supports methods to compute over two dozen measures of location, dispersion and distribution shape, including arithmetic, geometric and harmonic means and median (but not mode), variance, confidence intervals, histogram, and several inferential methods for linear and logistic regressions, t-tests, F-tests, and outlier detection. We show the accuracy of the proposed rigorous approaches via numerical comparisons between them and other bounding techniques like global optimisation, and with other traditional statistical methods for handling censored data.