Statistics for imprecise data

The key to enlarging the IP community

Marco De Angelis and Scott Ferson, Institute for Risk and Uncertainty, University of Liverpool, UK
Luke Green, Vivaldi Analytics, Stony Brook, New York, USA

How will the field of Imprecise Probabilities (IP) grow?

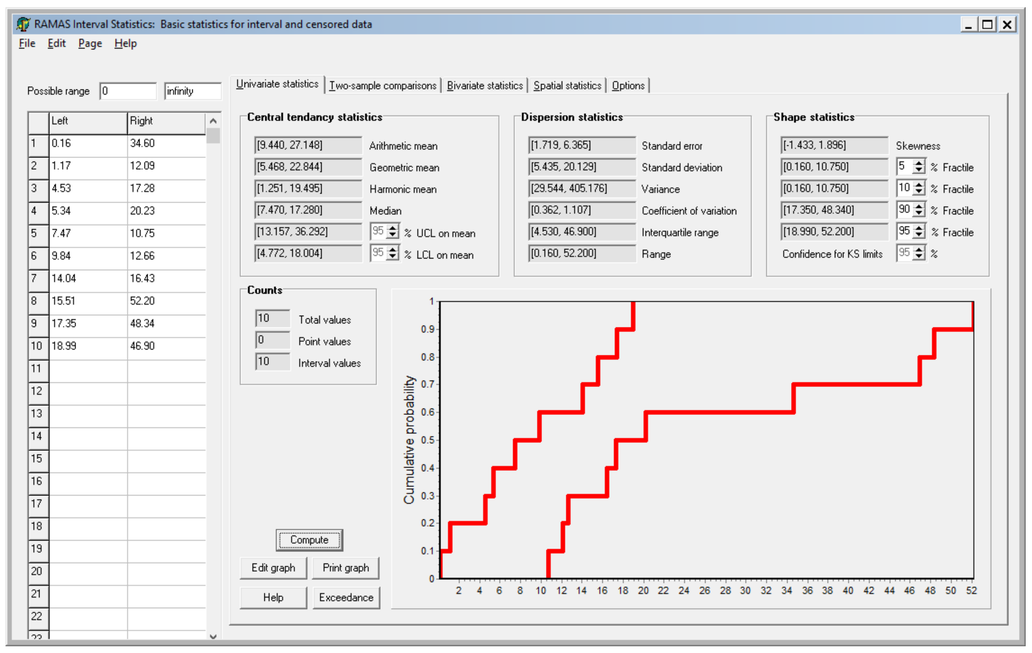
- IP is too complex for *almost all* users who see imprecision in their research.
- To grow, we need to recruit users who will apply IP in their routine work.
- We need convenient software that doesn't require special training in IP.

Statistics in the next century

- Statistics spent the last 100 years on methods for handling small sample sizes.
- But not all uncertainty in data has to do with limited sampling.
- Even in Big Data with huge sample sizes, imprecision can be substantial.
- People usually ignore imprecision because there isn't good software.

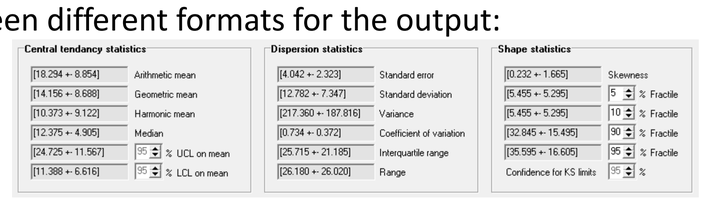
Solution

- New software offers convenient access to basic statistics for interval data.
- This software needs fewer assumptions than conventional methods handling imprecision, data censoring, missingness, or lack of independence.
- Computing many basic statistics for data sets with intervals is NP-hard.
- But many practically important special cases have efficient algorithms.
- Over two dozen measures of location, dispersion and shape, histograms, confidence intervals, and methods for regressions, *t*-tests, outliers, etc.
- C-code library, deployed on the cloud and in stand-alone software

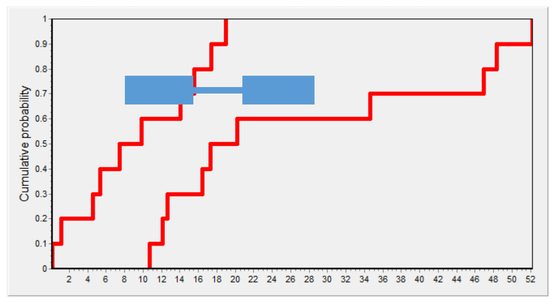


Users can choose between different formats for the output:

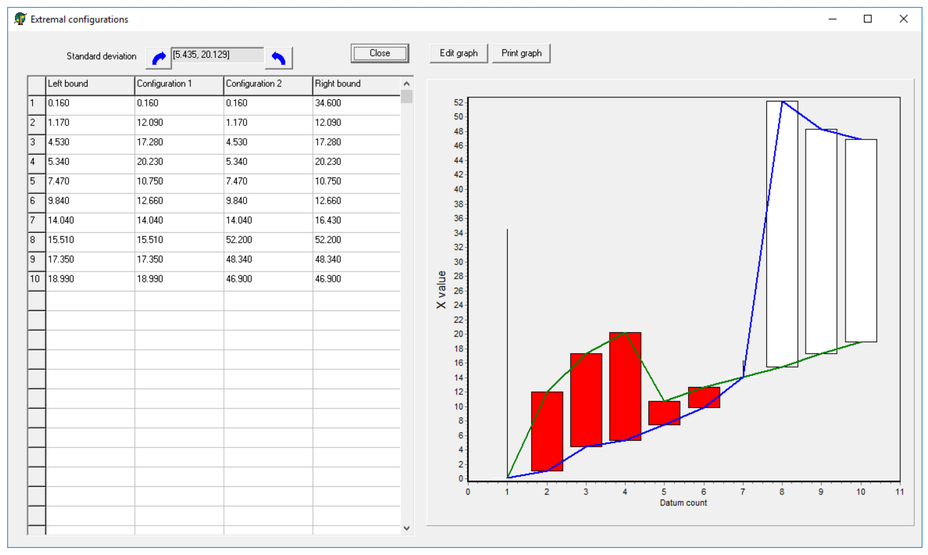
Interval [lo, hi]
Intervals [mid ± rad]
Significant digits
Scalars (ignore range)



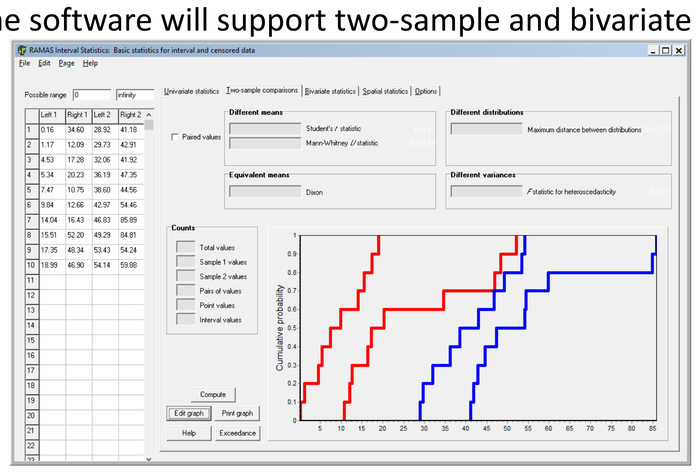
Clicking on any statistic displays it on the graph (e.g., standard deviation):

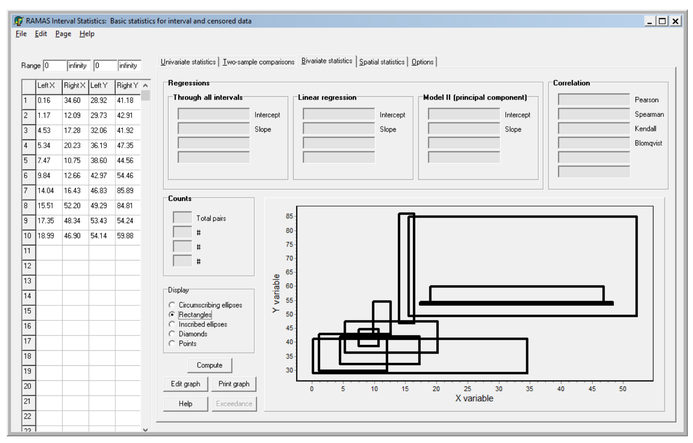


Double-clicking on a statistic shows the configuration of points within the respective intervals that underlie the extremal values of the statistic:



The software will support two-sample and bivariate statistics too:





When are data intervals?

Periodic observations

When did the fish in my aquarium die during the night?

Plus-or-minus measurement uncertainties

Coarse measurements, readings from digital devices

Non-detects and data censoring

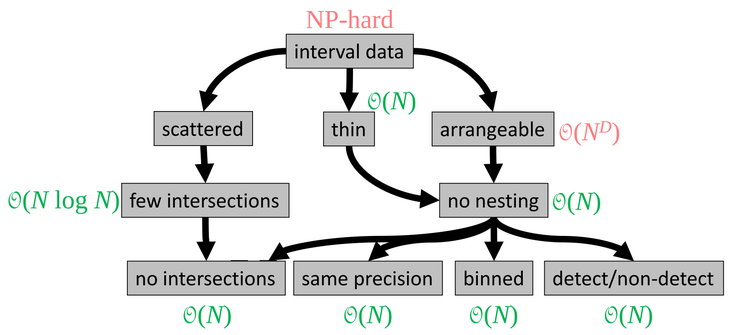
Chemical detection limits, subjects prematurely terminated

Privacy requirements

Epidemiological or medical information, census data

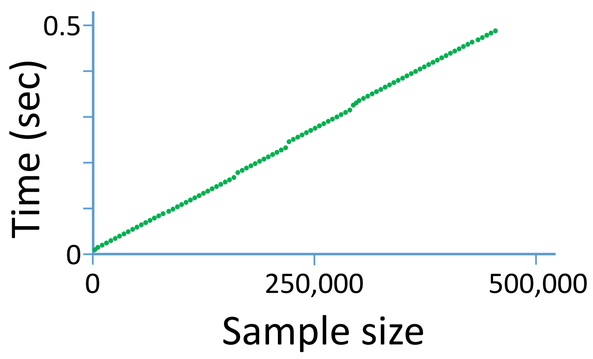
Computational difficulty

Computing variance (and many other statistics) for data sets with intervals is an NP-hard problem. Luckily, efficient algorithms are available for many special cases. The software detects whether a data set conforms to any of these special cases, and uses the most efficient algorithm available.



Computational speed

In practice, computation time for most data sets is not a serious barrier. For instance, in R, we can compute the variance for a data set of 400,000 intervals in under 0.5 seconds, a time comparable to that needed for a same-sized data set of point values.

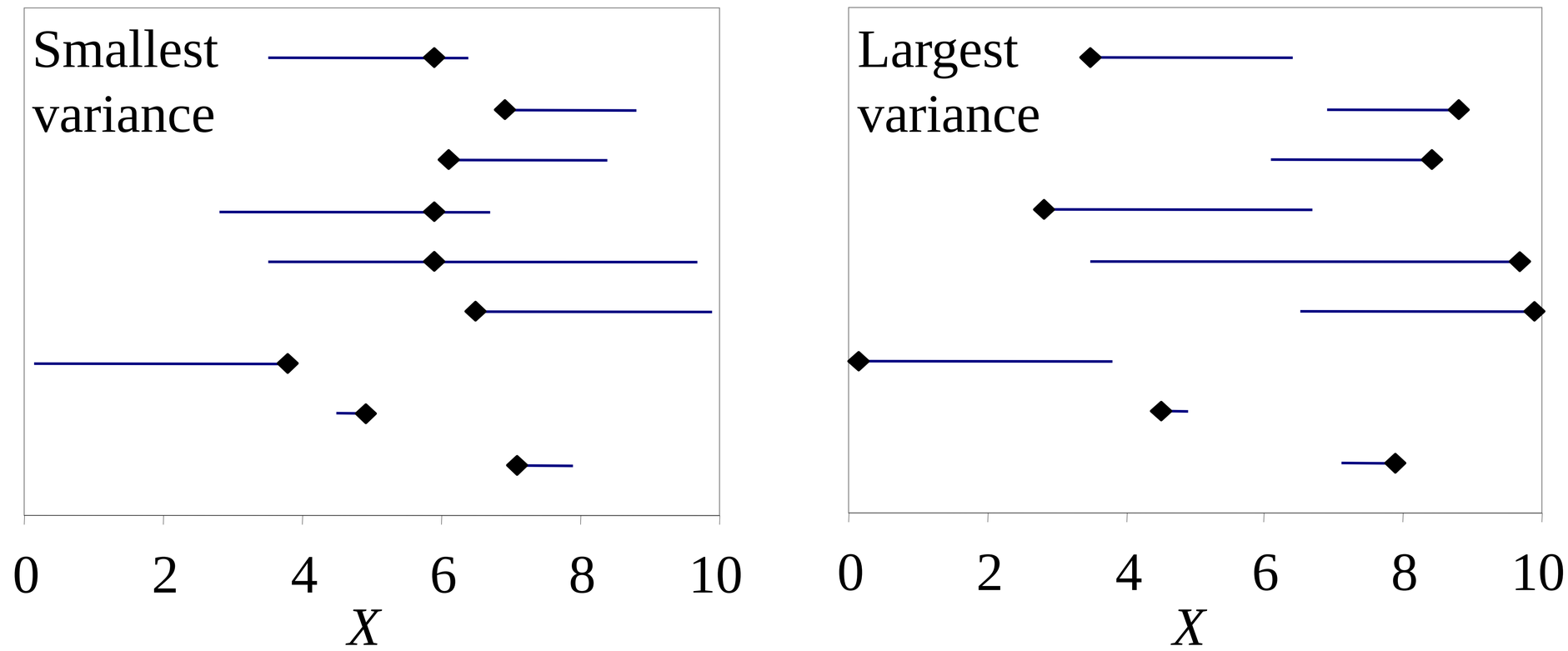


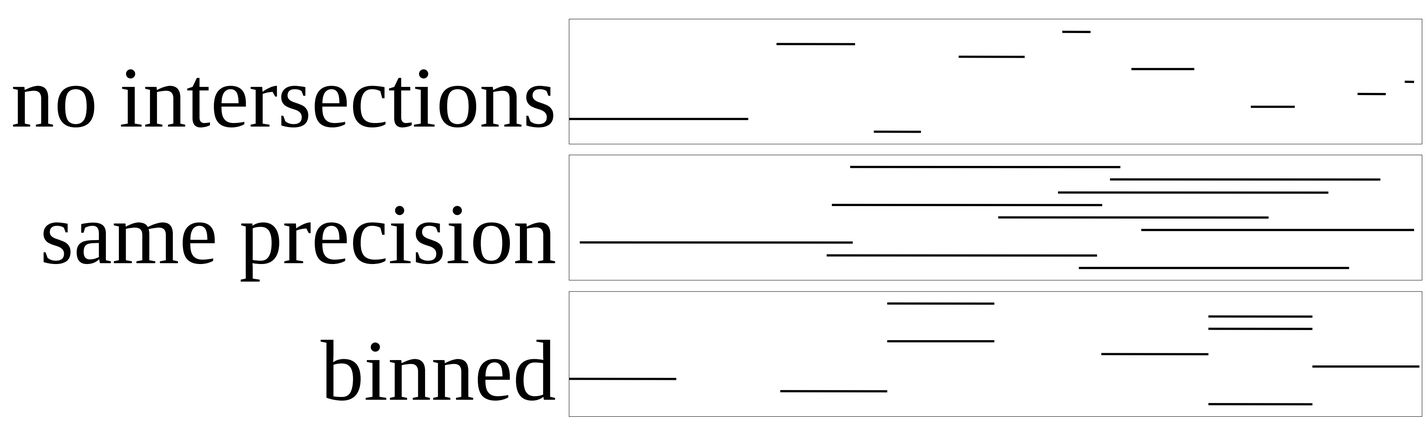
Conclusions

- Coherently combine data with different precisions
- Unify treatment for measurement imprecision (± parts), non-detects, arbitrary censoring, and missing data
- Straightforward statistics on data sets with intervals (with a few wrinkles) easily used in engineering calculations
- Convenient software (coming soon)
- Measurement imprecision means the error never goes to zero, *no matter how large the sample size*

Contact us

marco.de-angelis@liverpool.ac.uk
ferson@liv.ac.uk
luke@dataclimate.co





Unified treatment

- Measurement imprecision (± parts)
- Data censoring
- Non-detects
- Right-censoring
- Missing data

Wrinkles

- No mode statistic
- Most statistics are intervals
- Naïve interval methods grossly inflate uncertainty
- NP-hardness of the general case

Display options

Endpoint interval	[11.9, 12.3]	[10.1, 14.1]	[8.1, 16.1]
Plus-minus interval	[12.1 ± 0.2]	[12.1 ± 2]	[12.1 ± 4]
Significant digits	12	10	-
As scalar	12.1	12.1	12.1