Statistics for imprecise data The key to enlarging the IP community



Institute for Risk and Uncertainty



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

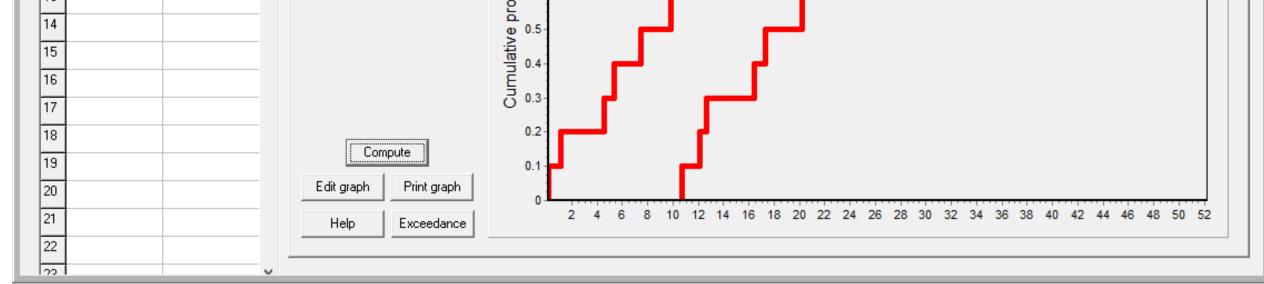
- Statistics spent the last 100 years on methods for handling small sample sizes.
- But not all uncertainty in data has to do with limited sampling.
 Even in Big Data with huge sample sizes, imprecision can be substantial.

🏹 R.4	MAS Interv	al Statistics: Basic	statistics f	or interval and censored data					<u> </u>			
ile	<u>E</u> dit <u>P</u> age	e <u>H</u> elp										
Pos	Possible range 0 Infinity											
	Left Right 🔨		^	Central tendancy statistics		Dispersion statistics		Shape statistics				
1	0.16	34.60	_	[9.440, 27.148] Arithm	netic mean	[1.719, 6.365]	Standard error	[-1.433, 1.896]	Skewness			
2	1.17	12.09	_	[5.468, 22.844] Geom	ietric mean	[5.435, 20.129]	Standard deviation	[0.160, 10.750]	5 🔹 % Fractile			
3	4.53	17.28	_	[1.251, 19.495] Harmo	onic mean	[29.544, 405.176]	Variance	[0.160, 10.750]	10 🔹 % Fractile			
4	5.34	20.23	_	[7.470, 17.280] Media	an	[0.362, 1.107]	Coefficient of variation	[17.350, 48.340]	90 🜩 % Fractile			
5	7.47	10.75	_	[13.157, 36.292] 95 🚍	% UCL on mean	[4.530, 46.900]	Interquartile range	[18.990, 52.200]	95 🚖 % Fractile			
6	9.84	12.66	_	[4.772, 18.004] 95 🚍	% LCL on mean	[0.160, 52.200]	Range	Confidence for KS limits	95 🛫 %			
7	14.04	16.43		Counts								
8	15.51	52.20			1							
9	17.35	48.34		10 Total values	: 0.9-							
10	18.99	46.90		0 Point values	0.8-							
11				10 Interval values	1							
12					Atilique 0.6	Г						
13					- 0.6-							

• People usually ignore imprecision because there isn't good software.

Solution

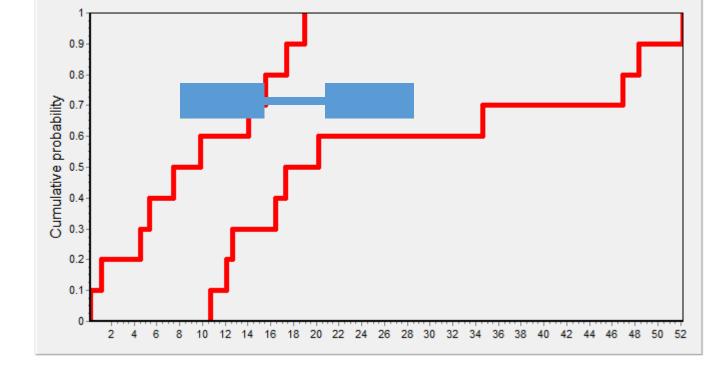
- New software offers convenient access to basic statistics for interval data.
- This software needs fewer assumptions than conventional methods handling imprecision, data censoring, missingness, or lack of independence.
- Computing many basic statistics for data sets with intervals is NP-hard.
- But 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:

Intorval [la bi]	Central tendancy statistics			Dispersion statistics			Shape statistics		
Interval [lo, hi]	[18.294 +- 8.854]	Arithmetic mean		[4.042 +- 2.323]	Standard error		[0.232 +- 1.665]	Skewness	
Intervals [mid ± rad]	[14.156 +- 8.688]	Geometric mean		[12.782 +- 7.347]	Standard deviation		[5.455 +- 5.295]	5 🚖 % Fractile	
Intervais [initia ± rau]	[10.373 +- 9.122]	Harmonic mean		[217.360 +- 187.816]	Variance		[5.455 +- 5.295]	10 🜩 % Fractile	
Significant digita	[12.375 +- 4.905]	Median		[0.734 +- 0.372]	Coefficient of variation		[32.845 +- 15.495]	90 🗢 % Fractile	
Significant digits	[24.725 +- 11.567]	95 🕏 % UCL on mean		[25.715 +- 21.185]	Interquartile range		[35.595 +- 16.605]	95 🚖 % Fractile	
Scalars (ignore range)	[11.388 + 6.616]	95 🗲 % LCL on mean		[26.180 +- 26.020]	Range		Confidence for KS limits	95 🗙 %	

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



Double-clicking on a statistic shows the configuration of points within

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

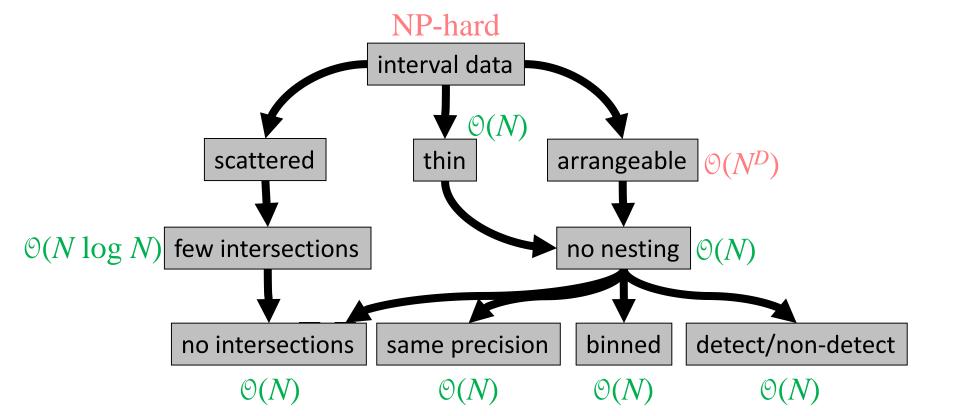
Non-detects and data censoring

Chemical detection limits, subjects prematurely terminated Privacy requirements

Epidemiological or medical information, census data

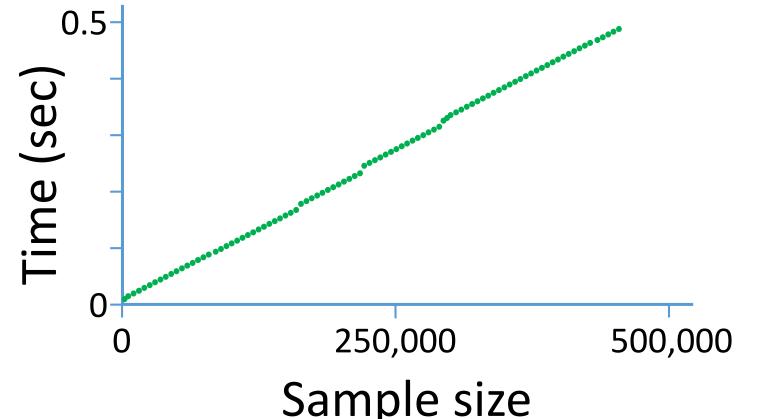
Computational difficulty

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.

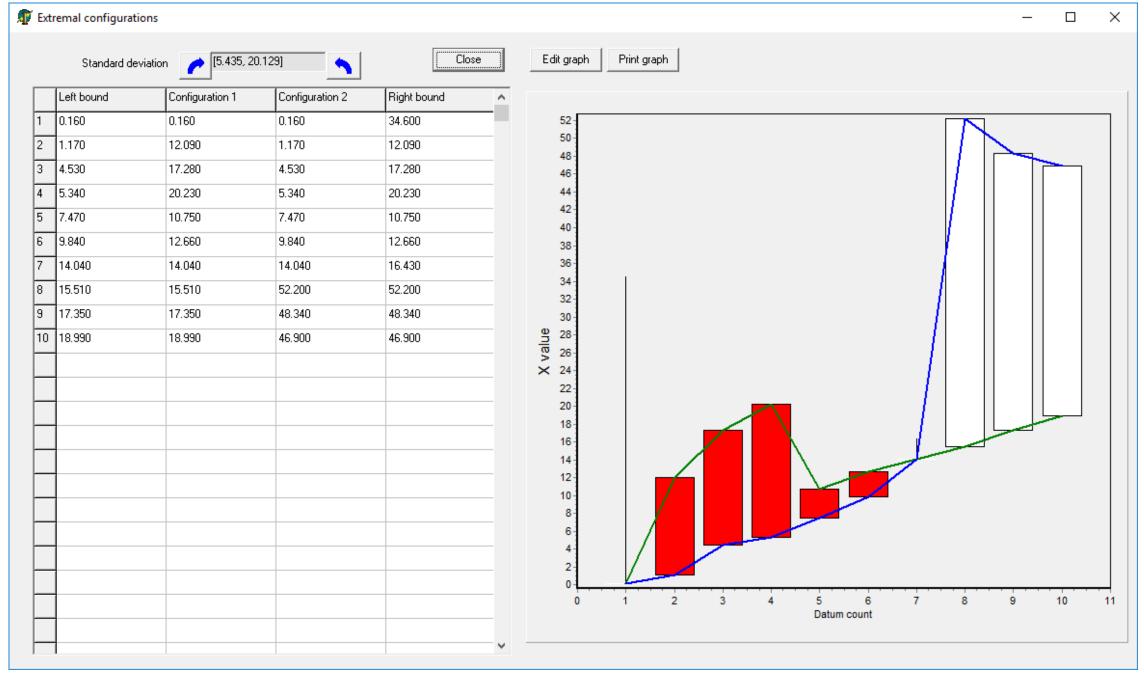


Computational speed

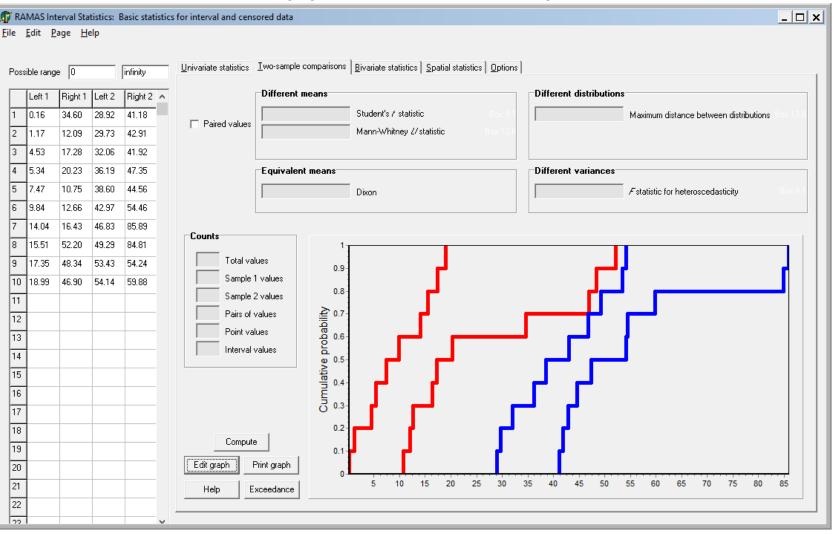
In practice, computation time for most data sets is not a serious barrier. For instance, in R, we can compute the variance for a data set of 400,000 intervals in under 0.5 seconds, a time comparable to that needed for a same-sized data set of point values.



the respective intervals that underlie the extremal values of the statistic:



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



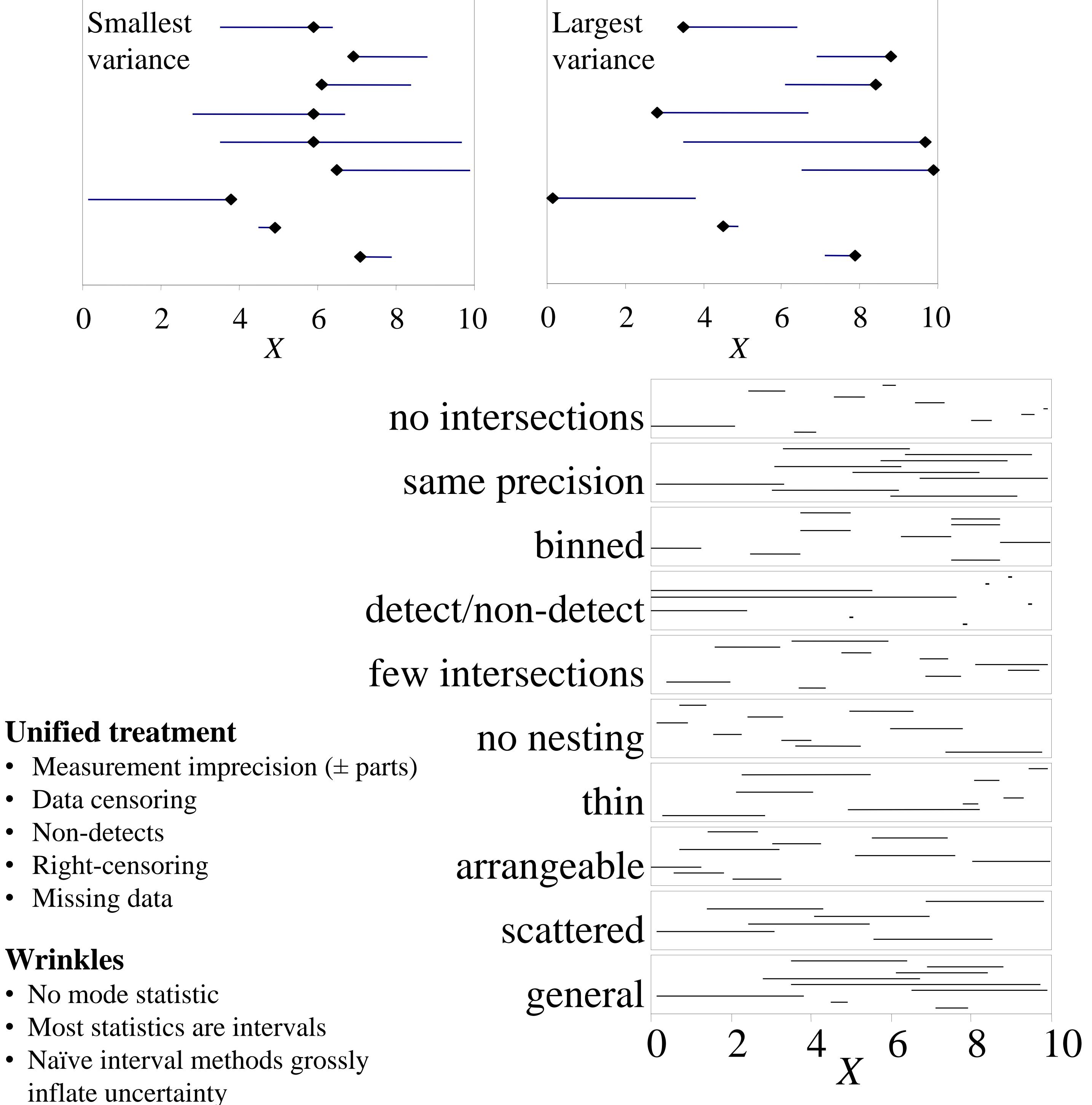
Conclusions

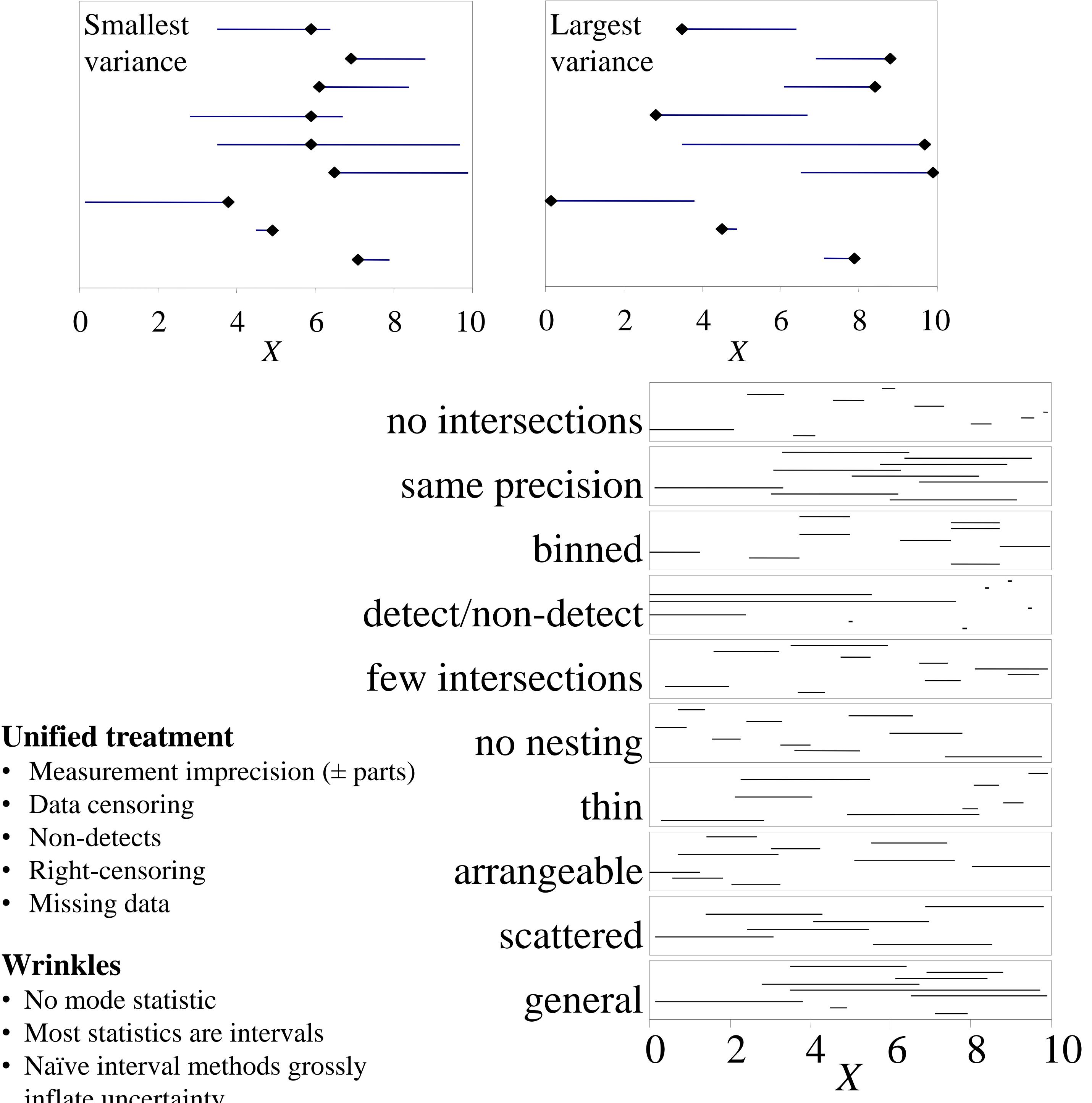
- Coherently combine data with different precisions
- Unify treatment for measurement imprecision (± parts), non-detects, arbitrary censoring, and missing data
- Straightforward statistics on data sets with intervals (with a few wrinkles) easily used in engineering calculations
- Convenient software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

na	ige 0	infinity	0	infinity	Univariate statistics	mple compariso	ons <u>B</u> ivariate statistics	<u>Spatial statistics C</u>	<u>]</u> ptions		
	Left X	Right X	Left Y	Right Y 🔺	Regressions					Correlation	
1	0.16	34.60	28.92	41.18	Through all intervals		Linear regression		Model II (principal component)		Pears
2	1.17	12.09	29.73	42.91		Intercept		Intercept	Intercept		Spea
3	4.53	17.28	32.06	41.92		Slope		Slope	Slope		Kend
4	5.34	20.23	36.19	47.35							Blom
5	7.47	10.75	38.60	44.56							
6	9.84	12.66	42.97	54.46							
7	14.04	16.43	46.83	85.89	_						
8	15.51	52.20	49.29	84.81	Counts						
9	17.35	48.34	53.43	54.24	Total pairs		85 -				
10	18.99	46.90	54.14	59.88	#		80 -				
11					#		75-				
12					#		70- 65-				
13						ple	60-				
14					Display	ari	55-				
15					C Circumscribing ellipses Fectangles	×	50-				
16					C Inscribed ellipses		45	╧╅┼┶╛			
17					C Diamonds C Points		40				
18							35-				
					Compute		30-				
19											

Contact us

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





• NP-hardness of the general case

Display options

Endpoint interval

Plus-minus interval

Significant digits As scalar

[11.9, 12.3] $[12.1 \pm 0.2]$

12

12.1

[10.1, 14.1] $[12.1 \pm 2]$

10

12.1

[8.1, 16.1] $[12.1 \pm 4]$

12.1