Responses and Questions for

"Systematics at LHC for event selection, discovery and limits", by Lukas Heinrich

and

"Precision measurements", by Sasha Glazov

Thomas R. Junk

Fermilab

Phystat-Systematics

November 1, 2021

Comments prepared from materials provided in advance – possibly different from final talks

Comments on Lukas Heinrich's talk, "Systematics at LHC for event selection, discovery and limits"

P. 15 of advance materials: "often unclear whether parameterized interpolation
would match the true response of the simulator at that point"

Even in multiple dimensions, you can knock a point out of the set of simulated samples and try to interpolate it from its neighbors. Difference between actual simulated and interpolated version is an "uncertainty". The gap thus induced is often bigger than the gaps between the simulated samples, so divide by two or even four (?).

To make the task easier, you might try off-axis interpolation studies at generator level.

Or see if your interpolation parameterization can replicate reweighted systematics.

More Comments on Interpolation

More subtle cases – what happens if there is a threshold of some sort in the range over which you are interpolating?

What is the statistical uncertainty on an interpolated prediction?
It relies on two or more simulated samples, and in some sense is an average of them (but the interpolation gets more like just one of the simulated samples the closer the parameter is to the simulation point), and thus ought to be less than the stat. uncertainty at an uninterpolated point.

Smoothing algorithms may use multiple samples to help interpolation.

How do we define "diminishing returns"?

Interpolation uncertainties like the ones Lukas describes can be eliminated by running more simulation at more model points.

But often this is not "worth it", which is why we're interpolating in the first place.

The usual definition of "worth it" is "does it reduce the total expected uncertainty, or shrink the expected confidence interval, or reduce the expected p-value."

What gets balanced on the other side is time, effort, staff.

What defines "diminishing returns"?

Collecting many small systematic uncertainties

Drop systematics below 0.5% - what if there are a lot of correlated ones?   (e.g. energy scale can be divided up among each tower of the calorimeter. Each one contributes little, but there's a correlated total energy scale uncertainty. We could game this rule and divide any "big" uncertainty into many droppable parts)

Can you collect the ones that have the same shape impact?

Or build a covariance matrix to cover a large number of unnamed systematics? (not my favorite solution, but if you have more nuisance parameters than you have bins, this might be more efficient).
This is done with the PPFX neutrino flux prediction package for hadronic prooduction uncertainties in the target. Many cross section uncertainties in GEANT4 – each bin of a cross section vs. energy plot for example.

Two-Point Uncertainties "Incompatible with MINUIT"

Some discrete nuisance paramteters (or parameters of interest!) are facts of life.

Example: the mass ordering in the neutrino sector.

One can minimize over continuous parameters for each
value of the discrete ones separately and take the minimum.

A combinatoric explosion ensues if you have several of these.

To be practical, one my want to trim the search space somehow.

Theory Uncertainties

Often, theorists will want their uncertainties treated in a different way than is customary for experimental uncertainties.

Often: uniform priors within a bounded region, zero outside.

Sometimes these are not credible, such as when a NLO calculation does not like within the quoted bounds of an LO calculation.

A small-order calculation may be missing enough terms to make it less sensitive to factorization and renomalization scale uncertainties.

Or: theorists may suggest using supremum p-values to be conservative

May be covered much more thoroughly in Frank Tackmann's talk

Split Outcomes

A null-hypothesis test p-value is the probability of getting a result as extreme or more extreme as the observed data, assuming the null hypothesis is true.

The "greater-than-or-equal" often has a substantial probability on the "or-equal-to" part, if the main experiment has a small number of expected events.

Including a subsidiary measurement in the calculation of a test statistic divides the main experiment's outcome into a variety of possibilities indexed by the subsidiary experiment's outcome.

Adding a systematic uncertainty and a subsidiary experiment can have make an analysis more sensitive than it is if they weren't there, just because outcomes are more finely divided.

Expected vs. Observed Uncertainties

An experiment with a downward fluctuation in the observed event count produces an estimate of the rate that has a smaller uncertainty than if it had an upward fluctuation.

Known issue when combining Poisson measurements using BLUE

But subsidiary measurements may have a similar effect. The "ur-prior" tells us what to expect from the subsidiary.

How to handle this in the Frequentist case? Need a sample from which to draw toys.

One-Sided Uncertainties

Roger Barlow wrote two papers, on asymmetric statistical uncertainties and asymmetric systematic uncertainties
https://arxiv.org/abs/physics/0406120
https://arxiv.org/abs/physics/0306138

One-sided uncertainties: "Best" estimate is on the edge of the interval?

Marginalization will produce an "average" answer that will be biased away from the best estimate.

Symmetrization is sometimes warranted, but sometimes not. When?

Is there a delta function of the prior at the best-guess value? If so, how big?

Does MINUIT get "stuck" on the edge of the allowed range of a nuisance parameter if there's a delta function there?

Bayesian marginalization calculations are less likely to get stuck if the likelihood function or the priors have discontinuities.

Comments on Sasha Glazov's talk on "Precision measurements"

- "Important to reduce the statistical component of systematic uncertainty"

Roger Barlow recommends running enough MC so that these are "chickenfeed"
arXiv:hep-ex/0207026

But how much is enough? We are often weighing effort and computational resources against physics sensitivity.

0.5% is "too big" for some measurements   (like *g*-2 of the muon or the electron)

Comments on Sasha Glazov's talk on "Precision measurements"

- When separating statistical and systematic uncertainties, in the Frequentist prescription, the nuisance parameters b_j are "fixed to their best-fit values" -- Is there a Bayesian prescription?

- Does this separation of statistical and systematic uncertainty meet the needs of the community?
    - Can we tell from it if a result is "systematics limited"? There could be a component with large stat. uncertainty and small systematics that will start contributing but only with a much larger dataset.
    - Does this distinction make sense for combinations, as some systematics have a statistical component

- Aside: ML community terminology: Aleatoric and Epistemic uncertainty. Not quite a map onto stat. and syst. uncertainty, as the use-case is different. Uncertainty in classification of one example. Training sample size and all other systs.

- How do we deal with unknown correlations? Published results often are missing correlation information.

Valassi, A., Chierici, R. Information and treatment of unknown correlations in the combination of measurements using the BLUE method. *Eur. Phys. J. C* **74,** 2717 (2014). https://doi.org/10.1140/epjc/s10052-014-2717-6

- Do one-sided uncertainties arise in precision measurements?

Comments on Sasha Glazov's talk on "Precision measurements"

How are analyses optimized?

Typically to minimize the expected uncertainty, the expected *p* value, or optimize the expected limit.

But sometimes no attention is paid to optimizing a combination with other analyses.

An end result, with a mixture of components: high stat. uncertainty, low syst. uncertainty, combined with high syst uncertainty low stat. uncertainty, can dilute information that would be needed in a combination of many similar analyses.

Ideally, the components are present for combination, not just the final result. The dominant component in a single result may not contribute much if it is systematics limited but many results have low-syst pieces with higher stat. uncertainties that eventually dominate the combined result.

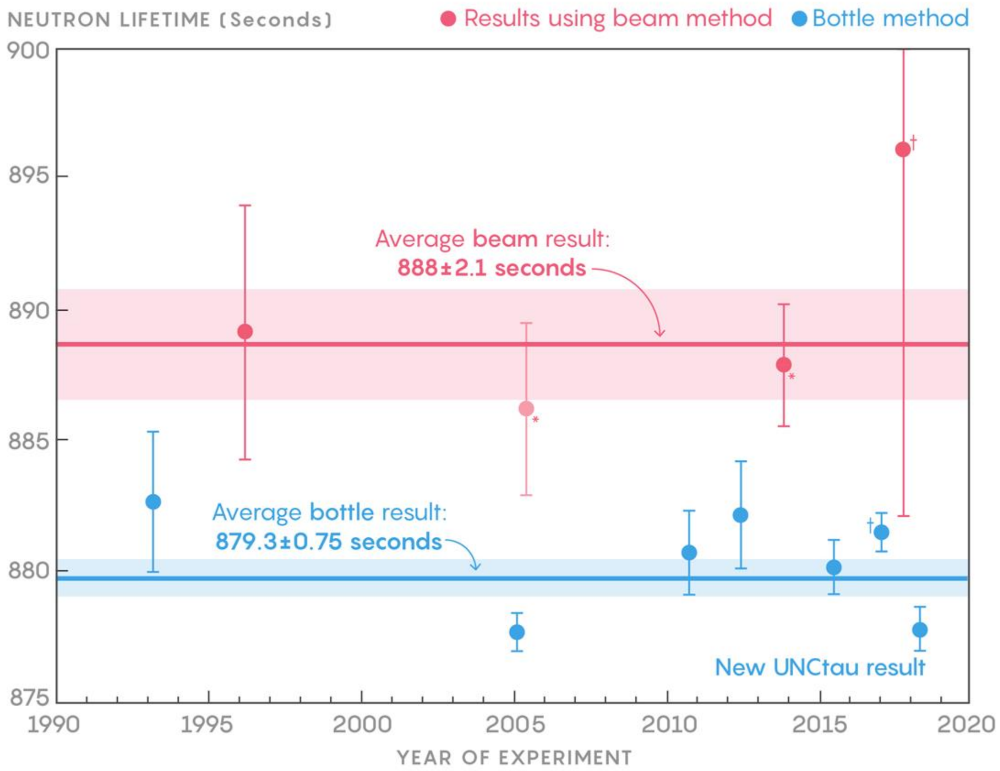
- How are results preserved for future re-use? More information than we're currently providing.

"PDG Method" of inflating uncertainties of large-chisquared data sets

- Goal is to get the best combined result
- But sometimes internal inconsistency might be telling us something about physics

Experiment comparison *p* value:
maximized over parameter
(neutron lifetime)
seems to make the most sense.

But what about nuisance parameters?
Do we also use the values that
minimize differences, even if the
priors tell us these values are not
credible?



*Nico result (2005) was superseded by an updated and improved result, Yue (2013);
†Preliminary results

https://www.quantamagazine.org/neutron-lifetime-puzzle-deepens-but-no-dark-matter-seen-20180213/

"unknown unknowns"