PHYSTAT-Systematics remote meeting, Nov 1st - 3rd and 10th, 2021:

INFORMATION FOR STATISTICIANS

In this note we want to explain a bit about the topic of this meeting, how the meeting is expected to work from the perspective of statisticians, and give a gloss on the list of topics which may come up in the meeting.

In order to prepare for this meeting we think it is important to try to bridge any language gaps between physicists and statisticians concerning statistical topics. The topic of the meeting is systematic uncertainties. Statisticians sometimes talk about systematic errors and think of them as biases in measurements. That idea is included here but the jargon 'systematics' goes well beyond that. Also, if a potential source exists, physicists try to correct for it; any uncertainty in correction is the corresponding systematic uncertainty.

SYSTEMATIC UNCERTAINTY AND NUISANCE PARAMETERS:

Statisticians are used to thinking about analyzing a single data set: writing down a statistical model for the data in some survey, study, or experiment and then using statistical procedures chosen to have desirable properties in this context. Often there will be many parameters (perhaps infinitely many) in the model. One common strategy in thinking about the best way to do some analysis is to split the collection of parameters into a low dimensional parameter of interest (often a scalar parameter) and the rest of the parameters; the latter are 'nuisance' parameters. Normally all these parameters are to be estimated using the single data set at hand; from here on, we are going to speak about the 'main experiment' which results in data which is to be analyzed to measure some physical parameter of interest.

Often, however, many of the parameters describing the physical model for the main experiment will have estimates which will not be derived solely from the data gathered in that experiment. Rather they may be measured in some or even many secondary experiments.

The paper by Joel Heinrich and Louis Lyons opens with a section which illustrates in a small experiment involving a pendulum a variety of types of systematic effects. A simple formula ($g = 4\pi^2 L/\tau^2$) connects the period τ and length L of the pendulum to the local acceleration g due to the Earth's gravitational field. Two systematic effects arise from potential bias in the measured length of the pendulum and in the period. Those effects might be corrected for in secondary experiments by calibrating the clock and the ruler. The uncertainties in those secondary measurements become systematic uncertanties for the main experiment. There are also further systematics arising from the fact that the idealised conditions under which the formula for g was derived might not be satisfied by the actual pendulum.

In general, a systematic effect is one which affects any of the measurements made – calibration errors in the measuring devices for instance or an error or an approximation in the theory which leads to the model being fitted. "Systematic uncertainty" is the

uncertainty in a measurement (the goal of the main experiment) induced by the uncertainty in the value of the parameter which describes this systematic effect.

THE STRUCTURE OF THE MEETING:

The timetable (soon to be available at the meeting's website) basically consists of:

- Three half-days with 2 sessions of 1.75 hours each. Most of the talks are by physicists. These talks will mainly describe how uncertainty is incorporated into the analysis of the experiment; some may also touch on how that uncertainty is estimated outside of the precise experiment being analyzed.
- Most of these 105 minute sessions have two talks, each of 30 minutes followed by general discussion, and then a 15 minute response from a statistician.
- There will be one session with a few 15 minute talks by statisticians.
- There will be two follow-up summary presentations by a physicist and a statistician (Sara Algeri) which will happen a week after the main part of the Workshop, on the afternoon of Nov 10th.

CHANCES FOR STATISTICIANS TO CONTRIBUTE

In live PHYSTAT meetings an important function of statisticians is just being there and being available for informal discussions with participants. Unfortunately this does not work for remote meetings, but we encourage statisticians to interact during or after the meeting with specific speakers, etc.

Other Statistician involvement includes the Introductory Statistics talk and the Concluding one. There are some 'Responses' by specific statisticians at the end of some sessions. We also have reserved a 1.75 hour session for short talks by statisticians. Please let us know if you would like to give such a talk.

Finally, we could consider having some further shorter (∼90 minute) follow-up meetings from time to time. These would include the possibility of talks by statisticians.

TOPICS LIKELY TO BE DISCUSSED AT THE MEETING

Here is our list of some things which are likely to be discussed at the meeting. The original list (red bullet points below) was written by Louis Lyons, then Richard Lockhart provided a Statistician's gloss on the terminology.

- Are systematics for the efficiency of event selection, parameter determination, discovery claims and upper limits dealt with differently?

  Discovery claims are essentially hypothesis testing problems, and are generally of great interest. Parameter determination and upper limits are two forms of confidence

interval problems, at least roughly. When a searched-for particle is not discovered, usually an upper limit is set on its possible production rate (assuming the particle exists), consistent with the data. In cases where the production rate is calculable in terms of the particle's unknown mass, some range of potentially possible masses can be excluded.

Typically 'parameters' will be estimated with attached errors:

Estimate ± Statistical Uncertainty ± Systematic Uncertainty,

where the systematic uncertainty includes contributions from all studied sources of systematic error. When, as is common, these different sources provide uncorrelated uncertainties the overall systematic uncertainty is measured by the usual formula for the SD: square root of the sum of the squares. Physicists often say the uncertainties are "added in quadrature".

Efficiency of event selection is another estimation problem. In a typical analysis the events of interest (signal) are a small/minute fraction of the total number of interactions recorded (mostly background). Typically a multivariate technique involving Machine Learning is used to separate a smaller sample with greatly reduced background, while still retaining a good fraction of the signal. A typical example involves estimating the rate at which some particle is produced in a collider. The estimate must be adjusted to compensate for events which had signal but were eliminated by the multivariate technique; the uncertainty in this inefficiency is a systematic.

- Some systematics have a statistical origin. Is it unimportant what they are called, so long as their correlations (with other possible measurements) are dealt with correctly?

  Some systematics arise from a parameter (or parameters) in the analysis of the current data set which was estimated in a secondary experiment. The secondary experiment is then used to produce a 'likelihood' for that parameter. This 'likelihood' then multiplies the likelihood for the main experiment. These nuisance parameters are then typically eliminated in the main experiment by profiling. This particular systematic for the main measurement thus has its origin in a statistical uncertainty in the subsidiary measurement. But even though this systematic has a statistical origin, it could still be correlated with the corresponding systematic in another experiment.

- Are there situations where profiling a likelihood is better than marginalising a posterior probability?

  The main issue here is the problem of summarizing the uncertainty in the parameter of interest in the main experiment. Multiplying the likelihood in the main experiment by these likelihoods from secondary experiments is like using the secondary likelihood as a prior. With these terms included should one eliminate the nuisance parameters

by integrating them out as a Bayesian might or using profile likelihood? This topic is one where statisticians must have lots to say: Bayes versus frequentist, frequency properties of Bayesian or partially Bayesian methods, and potential for bias resulting from profiling are all statistical topics. See also the discussion later of the Neyman-Scott problem.

- How do we deal with asymmetric uncertainties? e.g. How to combine several measurements with asymmetric uncertainties; calculating Goodness of Fit; etc.

  Symmetric uncertainties (symmetrically distributed uncertainties) often arise as the result of a Gaussian approximation. It is natural to summarize the uncertainty in terms of plus or minus some number of estimated standard errors; from such a summary confidence intervals with other desired coverages are easily computed. But sometimes the likelihood (or profile likelihood or marginal posterior) in the secondary experiment is not very Gaussian. For instance particle lifetimes are exponentially distributed; with small numbers of measurements this would give rise to a clearly skewed likelihood and an uncertainty which is not centred around the maximum likelihood estimate.

  In such a case one might move left or right far enough to lower twice the log-likelihood the same amount (say $z^2_{\alpha/2}$) in each direction. For profile likelihoods which are not very quadratic these departures might be different in the two directions. This would give an asymmetric interval which could be written in the form

  Estimate - DL to Estimate + DU.

  Here DL is the distance from the MLE to the lower limit of the likelihood interval and DU is the distance from the MLE to the upper limit of the interval. Is there any clear theory which would predict the amount of asymmetry in the profile likelihood so that the results of several such uncertainties could reasonably be combined? Or is the problem likely to be one of poor Gaussian approximations which have to be fixed in a case-by-case way?

  Another possible source of asymmetry could be from a non-linear effect of changes in the nuisance parameter on the estimate of the parameter of interest. There is likely scope for statisticians to be useful here.

  There is clear interest in the general topic of goodness-of-fit; see below for more.

- Do uncertainties in theory present special problems? And similarly when different theories make discrepant predictions.

  In trying to make predictions about the outcome of an experiment in which quantum mechanics plays a big role in describing the outcome one often makes use of approximations (expansions based on Feynman diagrams for example). The computations

must be truncated and approximated and this gives uncertainty in the predicted outcomes. This sort of uncertainty is not statistical. Can it be incorporated in any non-Bayesian way which is compelling? Do uncertainties like these mean that expert judgment is an unavoidable part of the uncertainty in the analysis? Is there a role for statisticians here? Combining expert opinion?

- Are there issues in situations with thousands of nuisance parameters? Can we run into problems as pointed out in the Neyman-Scott paper (Econometrica,1948)?

  There can be up to tens of thousands of parameters which are not measured in the main experiment but determined in secondary experiments. So the full likelihood for the whole model has a huge number of nuisance parameters. I think this problem, of trying to give general high dimensional nuisance parameter adjustments sounds like one statisticians are already working on. I don't think the vision is one where sparsity is a factor – the secondary experiments are thought, I believe, to make the full model identifiable. So one of the most popular high-dimensional strategies is likely not available. This might not be true about background modelling.

  In general, then, this is a question which should appeal to statisticians.

- The various issues that arise when Monte Carlo simulation is compared with the data, e.g. M.C./data discrepancies; weighted M.C. events; reweighting M.C. or generating new M.C. as parameter(s) of M.C. are changed; limited M.C. data.

  Perhaps it is useful to imagine that you are trying to estimate some parameter which is taken to have a real meaning: mass of a Higgs boson or some such thing. I think this is a 'parameter of interest' while the distribution of the background (events in the data set which don't include a Higgs boson) is clearly nuisance. The background itself will often be measured separately using a (very very large) Monte Carlo process which simulates the quantum mechanical nature of the particle interactions for producing the observed events, and also the detection processes of the entire detector. The kind of Monte Carlo statisticians usually do is called 'toy Monte Carlo'. I think there is a natural goodness-of-fit problem here: Is the real data consistent with signal plus Monte Carlo background?

  There is sometimes also Monte Carlo signal data (when the nature of the signal is predicted reasonably well by the theory being examined) and more goodness-of-fit problems suggest themselves. This is usually an easier problem than the problems arising from modelling the background.

  Reweighting is usually a version of importance sampling and probably a better name for the idea than ours.

- If the effect of changing a nuisance parameter on the parameter of interest is $x \pm s$ (where s is an uncertainty on the estimate x, perhaps arising from limited number of events in the M.C.), what do we take as the contribution of this nuisance parameter to the total systematics?

  Louis points out that all of the following ad hoc procedures (and more) have been suggested:

  $|x| + s$,
  $\sqrt{x^2 + s^2}$,
  zero if $|x|/s$ is less than 2, or $|x|$ otherwise,
  larger of $|x|$ and s,
  $\sqrt{x^2 - s^2}$.

  These various formulas appear to have quite different motivations suggesting lots of scope for statisticians to provide suggestions.

- Relative merits of Bayesian and Frequentist (likelihood) ways of incorporating nuisance parameters.

  This question seems pretty clear for statisticians once one has in mind a clear scope of meaning for 'nuisance parameters'. But the fundamental issues of the discipline, like what are we trying to optimize and what do our modelling assumptions amount to, are sure to remain. There are always statisticians willing to discuss 'foundations of statistics'.

- Validity of combining Bayes for nuisance parameters with frequentist methods for parameter of interest.

  This refers to a paper by Cousins and Highland where they use a frequentist approach for determining upper limits, while using a Bayesian method of averaging over nuisance parameters. (That is, it is not just using Frequentist coverage to check the behaviour of a Bayesian method.) The method is judged, empirically, to work well but this must be a topic of interest to statisticians.

- Problems in the shift method for estimating the effect of nuisance parameters. This includes OPAT = changing One Parameter At a Time. How does it compare with changing all parameters simultaneously? Do possible non-linearities affect the choice?

  In experimental design we try to assess lots of interaction effects as well as the main effects. The idea is, I think, that the estimate of the parameter of interest is roughly

linear in each of the nuisance parameters and you estimate the slope of this line by changing the one particular nuisance parameter you are thinking about without changing anything else and then see what happens. This sounds like an experimental design problem and the question is can you ignore second order effects including cross partial derivatives. Is there room for thinking about experimental design for assessing the combined effect of parameters whose nature is such that trying a new setting is computationally very expensive?

In principal, changing all the parameters together is just Monte Carlo for the distribution of a function of some variables $x, y, z, \cdots$. This discussion of quadratic terms and cross product terms suggests you are thinking of Taylor expansion of that function. These higher order terms are almost always not 0, exactly, but we routinely discard them as always. I don't think OPAT is much in use although my colleagues in 'Experimental Design' often mention it as a bad idea in an industrial experimentation context. They don't usually recommend randomly varying the parameter settings — rather they structure a set of parameter settings to permit estimation of linear terms (main effects in the jargon) and cross-product effects (two-way interactions). They often work in a framework where each parameter is tried at only 2 settings. They have, however, other methods when they are worried about quadratic and higher order shapes. Jargon: factorial experiments, fractional factorial experiments and lots more.

- The paper by Dauncey et al ([1408.6865] "Handling uncertainties in background shapes: the discrete profiling method" - arXiv.org e-Print archive) deals with discrete nuisance parameters (e.g. the functional form chosen for fitting a spectrum) in analogy with the way continuous nuisance parameters are profiled. Is this a known statistical procedure?

  This is a good question for statisticians.