April 20, 2023

Banff Challenge 3: Systematic Uncertainties

Tom Junk
Fermi National Accelerator Laboratory, Batavia, IL 60510 USA

The estimation and propagation of systematic uncertainties is a primary component of nearly every analysis of high-energy physics data. Deliverables from experiments include confidence intervals on model parameters and hypothesis test results in the form of p values. The sensitivity of the experiment – defined by the expected length of the confidence intervals it will produce, or the median expected p value under a specific alternate hypothesis, are affected by the lack of knowledge of the values of nuisance parameters. Even the number and type of nuisance parameters may not be known in a real experiment. Banff Challenge 3, described herein, is intended as a well-defined exercise in estimating and propagating systematic uncertainties in a simplified particle physics analysis. Simulated data distributions and model predictions in the form of random samples of pre-chosen distributions are provided. As is the case in a typical analysis, the simulated model sets are referred to as "Monte Carlo" while the simulated data sets are referred to as "data". The simulated Monte Carlo samples are labeled as "signal" or "background", while the simulated data samples are unlabeled. The simulated data and the simulated Monte Carlo samples differ in their underlying distributions. Features present in the data sample can be used to constrain these systematic differences. The object of the Challenge exercise is to produce 68% CL confidence intervals for the signal rate, including the effects of unknown nuisance parameters. The calculation of p values is not required. Proper coverage of the confidence intervals is required in order to win, and of the entries that have coverage, the winning entry will have the shortest expected intervals computed from the simulated data samples.

I. INTRODUCTION

Systematic uncertainties are an important component of nearly every analysis in high-energy physics (HEP). These uncertainties are included in results in order to attempt to cover biases between measured and true values of parameters. In analyses for which the end product is a confidence interval on a parameter of interest, these uncertainties most often increase the size of the expected and observed intervals. For many measurements, such as those in which the parameter being estimated is far away from a physical boundary and the uncertainties are relatively small, then the statistical and systematic uncertainties are often simply added in quadrature as an approximation to obtain the total uncertainty on a measurement, assuming Gaussian distributions. There are several methods for including the effects of systematic uncertainty in the calculation of p values [1], though this Challenge exercise does not test the ability to compute p values.

The differences between the known model assumptions and the unknown truth provides a difficult, ill-defined challenge. Systematic errors are usually parameterized with nuisance parameters. These parameters may be treated in analyses in either a Bayesian or a frequentist manner. In a Bayesian treatment, prior probability distributions are assumed for the nuisance parameters, and they are integrated out in the calculation of the posterior distribution. In the frequentist approach, the results of subsidiary measurements are used to constrain the values of the nuisance parameters, and the distributions of results of these subsidiary experiments are used in the calculation of the distributions of test statistics used to calculate p values and confidence intervals. In the case that no subsidiary measurement is available to constrain the value of a nuisance parameter but there is a distribution of belief in its values, such as from a set of theoretical calculations, a fictitious subsidiary measurement may be created for purposes of including the impact of that nuisance parameter [3].

One definition of them is "uncertainties in a measurement that do not dissipate as additional data are collected" [4]. It is often the case that the data collected by an experimental procedure can be divided into subsets that have different uses in extracting the measured values of the parameter(s). For example, a sought-after particle may have more than one decay mode, and event selection requirements can be optimized to select signal events of each decay mode, separating the data into discrete categories. The background rates in some of these samples may be larger than in others, or less well understood. It is often the case that rarer processes have smaller a priori uncertainties on their theoretical predictions than high-rate processes, especially if the rarer processes are governed by the electroweak interaction and the high-rate processes are governed by the strong interaction, which is notoriously difficult to model precisely. A combination of a statistically weak but systematically accurate measurement on a subsample of the data

(high variance, small bias) with a statistically strong but systematically weak subsample (small variance, large bias) will eventually favor the former as the data sample size grows. Control samples may also be statistically limited with small data set sizes making experimenters rely more heavily on theoretical predictions, while with more data, the statistical uncertainties shrink below the systematic uncertainties of the predictions. Both of these effects contribute in some analyses to allow the total uncertainty at large sample sizes to be smaller than the systematic uncertainty for small sample sizes.

Systematic uncertainties can be categorized in the following way, with many thanks to P. Sinervo [5], but slightly modified:

- The *good*: Nuisance parameters for which subsidiary experiments are available and well-understood frequency distributions can be modeled and/or posterior distributions are available which can be used as priors for the present analysis. Nuisance parameters that are shared between the main experiment and the subsidiary experiment are identified and central values adjusted for consistency. Additional nuisance parameters in the subsidiary experiment or in the extrapolation to the main experiment may weaken the effectiveness of this technique.
- The *bad*: Known sources of bias for which subsidiary experiments are unavailable. Calculations and guesses are used to estimate priors on the corresponding nuisance parameters. The extrapolation of the value of a "good" nuisance parameter from a subsidiary result to the main result may involve one or more "bad" parameters.
- The *ugly*: Sources of bias that are not considered, forgotten, or improperly dismissed. Many sources of uncertainty are in fact known to be small and can be dismissed based on simple calculations. But a large source of error that is not included in the uncertainty budget of an analysis will cause interval sizes and p values to be underestimated, producing undercoverage and embarrassing false discoveries [6].

There are three challenging tasks in the handling of systematic uncertainties. The first of these is the identification of all significant sources of bias. The second is the estimation of the magnitudes of the possible bias, and the construction of prior distributions for the nuisance parameters or the identification and use of subsidiary measurements to help constrain the values of the nuisance parameters. The third task is to include the effects of the nuisance parameters in the output products of the analysis. Banff Challenge 3 is intended to test participants' skills in performing these tasks in a simple, self-contained toy scenario. Enough information is given to help analyzers so that the estimation of uncertainties does not devolve into a mere guessing game. Sadly, real-world analyses often involve some guesswork.

II. THE CHALLENGE EXERCISE

All data files and this writeup are available at

`https://drive.google.com/drive/folders/1i2yDyiQo7wQOw0hGv2guwSPwAgIuCfdo`

A. Data and Simulation Samples

The exercise presented here is loosely based on an analysis of the rate of $W^{\pm}$ boson production at CDF that used two feature variables to separate the signal from the background. For a real-world example histogram, see Fig. 2 of Ref. [5]. These variables were the missing transverse energy and the amount of energy in a cone around a lepton candidate ("MET vs ISO"). More modern techniques exist for selecting $W^{\pm}$ events at hadron colliders and estimating the signal and background, but this one has the advantage of collecting all information in a single event sample with two feature variables. The distributions of the feature variables and their correlations are completely artificial in this Challenge exercise.

The two variables are called x and y in this exercise. The data consist of individual collisions in a particle-physics detector, each of which has a measured value of x and a measured value of y. A total of $n^i_{\rm obs}$ collisions are recorded in the $i^{\rm th}$ simulated data sample. Two processes contribute to the collision data – a "signal" process and a "background" process. The distribution of the features of the signal process is expected to be approximately independent in the x and y variables:

$$p_s(x,y) \approx f_s(x) g_s(y) \tag{1}$$

and the same is assumed to be true for the background process:

$$p_b(x,y) \approx f_b(x) g_b(y) \tag{2}$$

Violations of these factorization assumptions are to be included in the systematic uncertainty studies. Violations of the independence of x and y may be implemented practically by introducing one or more additional sources of signal and background, or equivalently, by choosing $p_b(x, y)$ simply not to factorize. A typical distribution of $p_s(x, y)$ is shown in Figure 1, and a typical distribution of $p_b(x, y)$ is shown in Figure 2. An example mixture of signal and background is shown in Figure 3. The goal of the Challenge exercise is to estimate the signal strength in each data set provided. Section II B provides more details of the specific deliverables to be provided by Challenge participants. The primary obstacle in this exercise is that the background production rate is unknown *a priori* and must be evaluated from the data. Indeed, it is chosen to have a different value from one Challenge data set to another. Complicating the rate extraction process is the fact that the shape of the background distribution is also not known perfectly.

The procedure Sinervo describes in Ref. [5] for dealing with the unknown background rate is to use an ABCD method with four regions defined in the two-dimensional feature space. Three of these regions are used to constrain the background rate in the fourth. Sinervo mentions that while the variables x and y are approximately uncorrelated, small, unknown correlations between x and y in the signal and also in the background probability distributions are sources of uncertainty on the measured signal rate. The procedure suggested in [5] for evaluating the systematic uncertainty is to adjust the boundaries of the A, B, C and D regions to see how robust the result is under this variation. This procedure is not without pitfalls, however. Roger Barlow says that varying selection requirements is a robustness check rather than an estimate of systematic uncertainty, and if anything, it only tells the analyzer about the distributions of events near the cuts and not deep within the A, B, C or D regions [7].

It is worth mentioning that eq. 4 in Sinervo's writeup [5] contains an erroneous term for the contribution of the background uncertainty to the total systematic uncertainty.

While the ABCD method has been used for measurements similar to that proposed here, Challenge participants are encouraged to explore other solutions and techniques.

Challenge participants are provided with simulated Monte Carlo data samples sampled from $p_s^{\mathrm{MC}_i}(x, y)$ and $p_b^{\mathrm{MC}_j}(x, y)$ for the i^{th} signal Monte Carlo sample and the j^{th} background Monte Carlo sample. The exact functional form of $p_s^{\mathrm{MC}_i}(x, y)$ and $p_b^{\mathrm{MC}_j}(x, y)$ is usually unknown to the analyzers as it is a cumbersome mixture of physics event generator models and detector simulation and reconstruction analysis chains. The Monte Carlo samples are all that is available. The indices i and j enumerate a discrete set of available Monte Carlo models. Typical examples of these include PYTHIA [8], HERWIG [9, 10], POWHEG [11–13], MCNLO [14], MADEVENT [15], and others. In fact, for simplicity, only one simulated alternative generator is explored in the data files provided, and it is not generated with a physics generator like the ones mentioned, but rather just a separate probability distribution. The simulated Monte Carlo samples are "labeled", in that the simulated collisions are known to be from the signal or background processes. In this Challenge exercise, it is intended that the background shape is more uncertain than the signal shape.

Files are provided as ASCII-formatted text files with two columns, separated by one or more spaces. The first column contains the x value for a particular collision, and the second contains y. Filenames are of the form `bc3_mc_bg_g`*n_syst*, where n is a generator index and *syst* is an indicator of what the systematic variation is. The values for *syst* can be `central` for the central-value signal, `npkp` for a $+1\sigma$ variation of the k^{th} nuisance parameter, and `npkm` for a -1σ variation of the k^{th} nuisance parameter. Signal Monte Carlo files have names of the form `bc3_mc_sig_g`*n_syst* with similar meanings. In this Challenge exercise, there are only two possible "generators" for the background Monte Carlo and one for the signal Monte Carlo. There is only one sample of alternate generator background. systematic variations are paired "down" and "up" variations of nuisance parameters. The nuisance parameters for the signal and background Monte Carlo sampls are *different* from each other and are not to be considered correlated. Parameter number 1 for the signal Monte Carlo is different from parameter number 1 for the background Monte Carlo. A list of the Monte Carlo samples is given in Table I.

A set of simulated "data" files is also provided, with names of the form `dataset`n`.dat`, where n is between 0 and 99, inclusive. The data samples consist of shuffled mixtures of signal and background collision samples. Quantum-mechanical interference between the signal and background is not simulated – the event samples are simple mixtures. The fraction of signal in the data will vary from sample to sample, although it will never be negative. As was the case for Banff Challenge 2, a fraction of the samples have zero true signal present. Other samples will have signals whose strengths are barely detectable, to test the ability of participants to create confidence intervals that are close to the boundary of zero signal strength. Still others will have signals that have significant strengths. The data files will not all have the same number of entries in them. The background rate is is varied from one data sample to another, so that Challenge participants cannot simply use the total number of data counts as a statistic for determining the signal rate.

The simulated Monte Carlo sets are meant to provide a sampling of possible shapes of the probability distributions for signal and background. The rate of background event production is to be considered not to be predicted by the models, nor is the signal rate, as it is the parameter of interest to be measured with the data. All of the signal and background Monte Carlo event samples therefore have the same number of events from one sample to another.

TABLE I: List of Simulated Monte Carlo samples provided with Banff Challenge 3's data sets

Filename	Meaning
bc3_mc_bg_g1_central.dat	central bg sample, generator 1
bc3_mc_bg_g1_np1p.dat	bg sample, n.p. 1 varied by $+1\sigma$
bc3_mc_bg_g1_np2m.dat	bg sample, n.p. 2 varied by -1σ
bc3_mc_bg_g1_np2p.dat	bg sample, n.p. 2 varied by $+1\sigma$
bc3_mc_bg_g1_np3m.dat	bg sample, n.p. 3 varied by -1σ
bc3_mc_bg_g1_np3p.dat	bg sample, n.p. 3 varied by $+1\sigma$
bc3_mc_bg_g2_central.dat	bg sample, generator 2
bc3_mc_sig_g1_central.dat	Central signal sample, generator 1
bc3_mc_sig_g1_np1m.dat	signal sample, n.p. 1 varied by -1σ
bc3_mc_sig_g1_np1p.dat	signal sample, n.p. 1 varied by $+1\sigma$
bc3_mc_sig_g1_np2m.dat	signal sample, n.p. 2 varied by -1σ
bc3_mc_sig_g1_np2p.dat	signal sample, n.p. 2 varied by $+1\sigma$
bc3_mc_sig_g1_np3m.dat	signal sample, n.p. 3 varied by -1σ
bc3_mc_sig_g1_np3p.dat	signal sample, n.p. 3 varied by $+1\sigma$

The simulated data on the other hand is drawn from similar distributions but not necessarily exactly ones represented by the Monte Carlo. This choice is meant to simulate all three kinds of systematic uncertainty. Furthermore, the distributions in the simulated data samples are to be considered statistically and systematically independent of one another – challenge participants should not use any property of one simulated data sample to help interpret any other simulated data sample.

Simulated data and Monte Carlo files are provided as a single tarfile, compressed with `bzip2`, called `bc3_challenge_datasets.bz2`. It contains two directories, `datasets` and `mcsets_labeled`. To unpack it, use the command `tar -jxf bc3_challenge_datasets.bz2` on most Linux and macOS systems.

B. Deliverables

Challenge participants are expected to provide, for each data sample, a 68% confidence interval for the signal rate in units of numbers of events. A real measurement of a cross section would divide the signal rate by the flux (or equivalently, the integrated luminosity) and the detection and selection efficiencies – these are not intended to be part of the exercise. In the interests of reducing the required volume of data to be transferred to challenge participants, no p-values are required.

Results are to be provided as a text file, with columns as follows: sample number, best-fit signal rate (in events), 68% CL interval lower edge for the signal rate (in events), and the 68% CL interval upper edge (in events).

A brief writeup is also to be provided along with each submission.

III. CRITERIA FOR WINNING

In order to win the competition, an entry's confidence intervals must cover the unknown true value 95% of the time. Among the entries, the one with the smallest average confidence interval length among the provided datasets will be chosen as the winner.

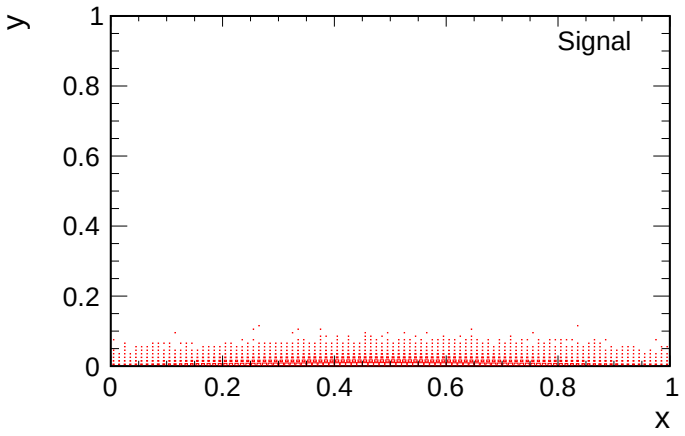

FIG. 1: An example distribution of the signal process in the (x, y) feature plane.

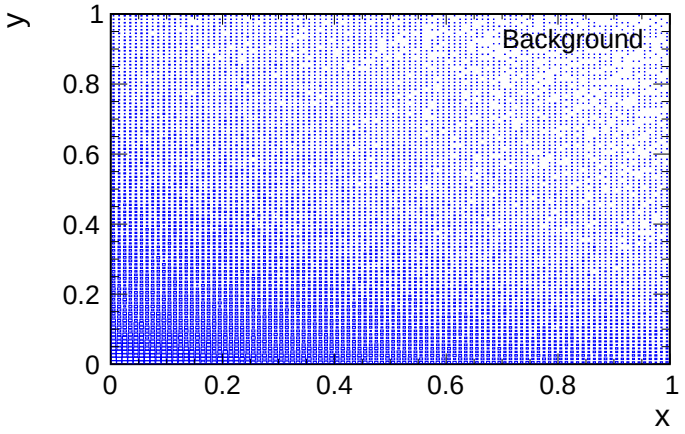

FIG. 2: An example distribution of the background process in the (x, y) feature plane.

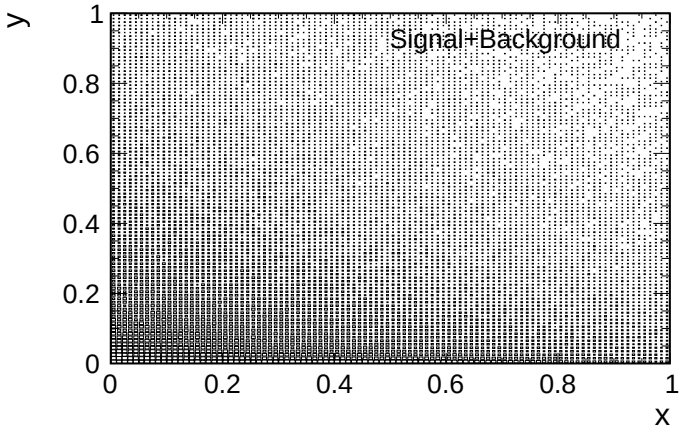

FIG. 3: An example mixture of unlabeled signal and background events in the ratio of 1:10 in the (x, y) feature plane.

[1] L. Demortier, "P values and nuisance parameters," contribution to PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, L. Lyons, H. B. Prosper and A. De Roeck, doi:10.5170/CERN-2008-001

[2] Baker, R.D. and Jackson, D. (2013), Meta-analysis inside and outside particle physics: two traditions that should converge?. Res. Syn. Meth., 4: 109-124. https://doi.org/10.1002/jrsm.1065

[3] [ATLAS, CMS and LHC Higgs Combination Group], CMS-NOTE-2011-005.

[4] D. Van Dyk, "Statistician's Overview of Systematics," presentation given at Phystat-Systematics, 2021. https://indico.cern.ch/event/1051224/timetable/

[5] P. Sinervo, eConf **C030908**, TUAT004 (2003) PHYSTAT-2003-TUAT004.

[6] T. R. Junk and L. Lyons, doi:10.1162/99608f92.250f995b [arXiv:2009.06864 [physics.data-an]].

[7] R. Barlow, [arXiv:hep-ex/0207026 [hep-ex]].

[8] C. Bierlich, S. Chakraborty, N. Desai, L. Gellersen, I. Helenius, P. Ilten, L. Lönnblad, S. Mrenna, S. Prestel and C. T. Preuss, *et al.* doi:10.21468/SciPostPhysCodeb.8 [arXiv:2203.11601 [hep-ph]].

[9] M. Bahr, S. Gieseke, M. A. Gigg, D. Grellscheid, K. Hamilton, O. Latunde-Dada, S. Platzer, P. Richardson, M. H. Seymour and A. Sherstnev, *et al.* Eur. Phys. J. C **58**, 639-707 (2008) doi:10.1140/epjc/s10052-008-0798-9 [arXiv:0803.0883 [hep-ph]].

[10] J. Bellm, S. Gieseke, D. Grellscheid, S. Plätzer, M. Rauch, C. Reuschle, P. Richardson, P. Schichtel, M. H. Seymour and A. Siódmok, *et al.* Eur. Phys. J. C **76**, no.4, 196 (2016) doi:10.1140/epjc/s10052-016-4018-8 [arXiv:1512.01178 [hep-ph]].

[11] P. Nason, JHEP **11**, 040 (2004) doi:10.1088/1126-6708/2004/11/040 [arXiv:hep-ph/0409146 [hep-ph]].

[12] S. Frixione, P. Nason and C. Oleari, JHEP **11**, 070 (2007) doi:10.1088/1126-6708/2007/11/070 [arXiv:0709.2092 [hep-ph]].

[13] S. Alioli, P. Nason, C. Oleari and E. Re, JHEP **06**, 043 (2010) doi:10.1007/JHEP06(2010)043 [arXiv:1002.2581 [hep-ph]].

[14] S. Frixione and B. R. Webber, JHEP **06**, 029 (2002) doi:10.1088/1126-6708/2002/06/029 [arXiv:hep-ph/0204244 [hep-ph]].

[15] F. Maltoni and T. Stelzer, JHEP **02**, 027 (2003) doi:10.1088/1126-6708/2003/02/027 [arXiv:hep-ph/0208156 [hep-ph]].