FERMILAB-CONF-11-363-E

Banff Challenge 2

Thomas R. Junk
Fermi National Accelerator Laboratory

Abstract

Experimental particle physics collaborations constantly seek newer and better ideas for improving the sensitivity of their searches for new particles and phenomena. Statistical techniques are the last step in interpreting the results of an experiment; they are used to make discoveries (hypothesis testing), and to measure parameters (point estimation). They are also used in the first step – experiment and analysis design. Banff Challenge 2 asks participants to test their methods of discovering hidden signals in simulated datasets and of measuring the properties of these signals. The Challenge problems are described, and the performances of the submitted entries is summarized, for datasets with and without simulated signals present.

1 Introduction

Experimental particle physicists are becoming more aware as time goes on of statistical techniques that have been developed in the context of other fields over the years, and are interested in new research as well, in order to maximize the usefulness of their experiments. To that end, an ongoing dialogue between physicists and statisticians has been very fruitful, providing useful benefits to both parties. Particle physicists gain knowledge of established and new techniques, and statisticians can explore their techniques with particle physics data. Unfortunately, the direct use of experimental collider data requires permission from a usually very large collaboration, and a substantial investment in understanding (and possibly improving) the modeling of imperfect detectors and imperfectly known physics processes. In order to simplify the process and allow as many people to participate as possible, well-defined "Challenge" problems have been created so that simulated data may be freely exchanged and results compared. This model of collaboration between statisticians and particle physicists was very successful in the experience of the first Banff Challenge [1], which explored the case of setting one-sided bounds on new physics processes in a Poisson counting experiment in which the background rate is constrained by an auxiliary Poisson counting experiment. We seek with Banff Challenge 2 to test methods of discovery and measurement.

2 Banff Challenge 2 Problems

Two problems were posed. Each has different features that illustrate some of the challenges faced by experimentalists when analyzing the data. In High-Energy Physics (HEP) language, the first problem consists of seeking a mass bump on top of an exponentially falling background. In the language of statisticians, the data are generated by a marked Poisson process – x is a quantity measured on each selected collision event, and is called a "mark". The signal and background distributions are given parametrically in this problem to simplify the treatment and to make the problem more accessible. The position of a localized excess is not specified in advance, although only one bump at a time is allowed to be present in this problem. Because the bump may be anywhere within the range of x provided, the issue of multiple testing, also called the "Look-Elsewhere Effect" (LEE), arises [2]. Participants were asked to measure the peak position and rate. The second problem asks participants to address a common analysis situation. The signal and background predictions against which the data are tested are provided by a Monte Carlo simulation and not an analytic parameterization. No LEE is present in the second problem.

In both problems, both the null and the test hypotheses are compound hypotheses – the background rates are subject to systematic uncertainty. In real HEP problems, the rates and shapes of the signals and multiple sources of background are uncertain, and these parameters are constrained by auxiliary data and/or approximate theoretical predictions.

In both problems, the task is to identify those simulated datasets that have signals injected in them and to measure the parameters of the signals. The Type-I error rate, that is, the rate at which evidence is claimed in the case that a signal is absent, should be no larger than 1%. Typically this is what is meant in a HEP experiment by "significance", that evidence is claimed by a method with a specified Type-I error rate. Participants then should optimize the power of their tests, that is, to claim evidence for a signal on a many simulated datasets that actually do contain a signal, while keeping the Type-I error rate within its bound. The correct evidence rate is $1 - \beta$, where β is the Type-II error rate, and this rate depends on the details of the signal injected, such as its strength and its position in the distribution of the marks. Participants were in addition asked to calculate their correct evidence rates for a small number of signal hypothesis choices, which were also among the choices tested in the blind samples. The estimation of the power of the test by the participant models an important ingredient in a HEP experiment. Physicists who propose building an experiment or conduct a specific analysis with an existing apparatus must justify their efforts to their colleagues and to their funders. They must provide a convincing argument that their experiment can provide a result that is interesting – that it is possible to find evidence for the sought-after new particle or process, or to exclude it if it is not present. Such estimates are used in decisions that allocate resources among experiments, and it is important that these estimates are neither underestimates nor overestimates. Banff Challenge 2 provides a blind way to check the methods used to arrive at power and coverage estimates.

The criteria for "winning" the competition among participants were not spelled out explicitly when the challenge problems were posed. It was made clear that a Type-I error rate of 1% or less is important to satisfy, and the best power satisfying the Type-I error rate requirement is a natural ordering to rank methods. Nonetheless, since we would like to test also the ability to estimate power properly, and because HEP experimentalists usually only have the estimated powers to use to rank methods instead of blind tests, it becomes natural to rank the methods based on their estimated powers, provided that the measured powers are not measurably below the estimated powers.

The problems are discussed in more detail below. The Challenge problems and simulated datasets may be found on the author's web site [3]. A summary of the submissions is available as well [4].

2.1 Problem 1

For this problem, the simulated data samples are drawn from the following density functions. Statisticians prefer the term "intensity" in order to indicate that they are not normalized to unit area. The background intensity function is

$$B(x) = Ae^{-Cx} \tag{1}$$

where x is the mark of the event. The domain of x is restricted to be between 0 and 1. We choose the values $A = 10000 \pm 1000$ and $C = 10.0 \pm 0$. The background rate parameter A is drawn from a truncated Gaussian distribution of width 1000, truncated so that $A \geq 0$. The signal intensity function is

$$S(x) = De^{-(x-E)^2/2\sigma^2} \tag{2}$$

The problem statement specifies that $D \geq 0$ and that $\sigma = 0.03$ and $0 < E < 1$ in the generation of simulated datasets. Two example distributions are shown in Figure 1, one for a signal injected near the upper end of the range of E, and one generated near the lower end of the range.

The Challenge datasets for Problem 1 were generated randomly according to the distribution $B(x) + S(x)$. There are 24 different subsets of simulated pseudoexperiments, corresponding to different choices of D and E, and these are listed in Table 1. The numerical choices were governed by

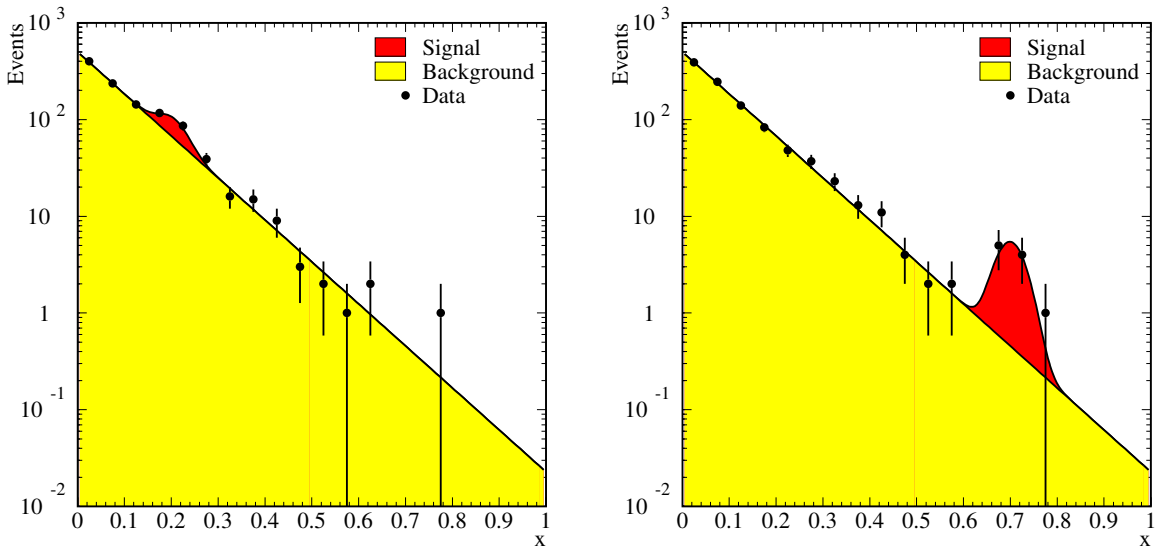

Fig. 1: Two example distributions for Problem 1. The simulated data are displayed in binned histograms, indicated by points with error bars drawn as $\sqrt{n_i}$ where n_i is count of simulated data events with marks that fall in that bin. Analytic functions for the signal and background intensities are also shown. Left: a signal injected with $E = 0.2$, and Right: a signal injected with $E = 0.7$.

the desire to have a correct-discovery rate that can be measured accurately with a limited number of repetitions, and thus should not be too close to 0% or 100%, and that we would like to test regimes in which large numbers of data events are needed for discovery and regimes in which a single mark or two can make all the difference. Signals with large values of D and small values of E have an approximately Gaussian distribution of the signal yield, while signals with small values of D and large values of E are very sensitive to the Poisson nature of the data in sparsely populated areas of the distribution.

The parameters D and E are parameters of interest and are not affected by unknown values of nuisance parameters, of which there is only one in this problem, a simplification compared to a real experiment. Similarly, the location of a peak does not always correspond to the true value of the mass of a new particle, although the significance of a peak should not be affected by the uncertainty in the relationship between the measured peak position and the underlying process that makes events in the peak. Similarly, since the significance of a peak that is found depends on the comparison of the data with the prediction of the null hypothesis, uncertainties in the probability of the detector to detect a signal event and for the analysis technique to select it should have little impact on the significance of a peak that is found, although these effects do affect the expected sensitivity, signal rate measurements, and limits.

The background parameter A was chosen for each simulated dataset from its prior distribution, a Gaussian centered on 10000 with a width of 1000. An integer n_b was then drawn from a Poisson distribution whose mean is the total background integral from $x = 0$ to 1 using the randomly selected value of A. Then n_b marks x were generated from the exponential distribution $B(x)$. A similar procedure was followed for generating marks for the signal component, according to $S(x)$. The marks were then shuffled and written out to the Challenge dataset file. Simulated datasets from the 24 categories were also shuffled so that no clue to the injected values would be provided by either the ordering of the datasets or of the marks within a dataset.

Three standard signal models were chosen for the purpose of asking participants to estimate their

correct-discovery fraction, and correspond to categories 4, 9, and 21 in Table 1. The correct-discovery fractions were measured on the corresponding challenge datasets and compared with the participants' estimates.

The presence of a nuisance parameter in the null and test hypotheses complicates the definition of the Type-I error rate. One approach is to evaluate the Type-I error rate as a function of the true value of the unknown nuisance parameter(s). Another approach is to evaluate the Type-I error rate in the prior-predictive ensemble whose generation is described above. A third is to quote the largest Type-I error rate for a fixed range of values of the nuisance parameters. The ideal that a method should cover for all values of the nuisance parameter requires a specification of what is meant by "all". The approach here is to quote the error rate and the correct-discovery rates using the prior-predictive ensemble, although this is not the only valid definition. A method which has a Type-I error rate which is larger than the stated value, which is usually written in a high-energy physics publication as a confidence level or a significance level, is said to undercover and is unlikely to pass collaboration review.

A feature of Challenge Problems 1 and 2 is that signal rate intervals were requested only in the case that evidence is claimed, and the problem statement asks for zero to be reported if evidence is not claimed. These instructions reflect a flip-flopping procedure which is very commonly used in HEP. If a collaboration measures the mass of a new particle but does not claim evidence for the new particle, the result may be easily misconstrued. Coverage for signal rate and peak position measurements were only computed on the subset of simulated datasets on which evidence is claimed.

Not quoting the measured signal yield in simulated datasets for which evidence is not claimed biases upwards the measured signal yields and the intervals containing them. A simple example is the null hypothesis – the true signal rate is zero in null hypothesis simulated datasets, but in 1% of them, a method that is performing well should claim evidence for a signal. Even if the set of intervals for the signal rate cover properly for a method, selecting this sample of them will in general not have proper coverage. This is true to a lesser extent for test hypotheses with true signals present.

A final feature of Problem 1 is that at most one signal is present, at a single value of E. In a real experiment in which the signal is *a priori* unknown, there may be more than one signal present. Since most methods fit for the background rate in the process of testing for the signal, a second signal (or more) will change the background fit. One may legitimately ask whether all of the events are signal events from a broad spectrum of multiple signals, and this is where some theoretical input and auxiliary information from other experiments is needed to constrain the background prediction. For this problem, we treat the presence of at most one signal as auxiliary *a priori* information. The Challenge datasets were generated with no more than one signal in each.

2.2 Problem 2

Unlike Problem 1, Problem 2 parameterizes the predictions of the signal and background yields using finite samples of Monte Carlo. In a real HEP experiment, samples of collider data from control regions are sometimes used instead. From a statistical standpoint, these are very similar and are treated identically. Often there is an extrapolation uncertainty associated with using a different sample of data which pass different selection requirements, and which are used to predict the background in the sample passing the signal requirements. Monte Carlos are similarly fraught with uncertainty in their predictions, and these uncertainties are parameterized with nuisance parameters which are simplified in this Challenge problem to two – one for each background governing its rate.

The simulated datasets and the Monte Carlo samples were generated from smooth distributions for the marks. The distribution of the marks for Background 1 is given by

$$x = \min(1.0, 1.4y^{2.74}e^{-y/3}), \tag{3}$$

where y is uniformly distributed on the interval $(0, 1]$. Background 2 was generated with a uniform

Table 1: Problem 1 Challenge dataset categories, listing the input values of E and D, the signal peak position and the signal rate parameters, respectively. The first category is the null hypothesis. For the categories marked with a "*", the participants were asked to compute their expected correct-discovery rates. The column headed $n_{\rm rep}$ lists how many simulated datasets were supplied for each of the categories.

Category	$E_{\rm input}$ (location)	$D_{\rm input}$ (intensity)	$n_{\rm rep}$
1	—	0.00	15400
2	0.50	83.78	200
3	0.38	265.96	200
4*	0.10	1010.65	200
5	0.10	478.73	200
6	0.66	66.49	200
7	0.78	39.89	200
8	0.10	744.69	200
9*	0.50	136.97	200
10	0.90	15.29	200
11	0.50	190.16	200
12	0.14	664.90	200
13	0.50	163.57	200
14	0.38	531.92	200
15	0.14	1196.83	200
16	0.50	110.37	200
17	0.10	1276.62	200
18	0.90	20.61	200
19	0.66	132.98	200
20	0.90	12.63	200
21*	0.90	17.95	200
22	0.90	23.27	200
23	0.78	79.79	200
24	0.10	1542.58	200

distribution. The signal distribution was generated using

$$x = z^{0.21}, \tag{4}$$

where z is uniformly distributed on the interval $(0, 1]$. The Challenge problem hid these underlying parameterizations, and gave samples of 5000 simulated marks for each of the three processes – signal, Background 1, and Background 2. The *a priori* yields given to the participants, which would be obtained from auxiliary experiments or theoretical predictions, are 900 ± 90 events for Background 1 and 100 ± 100 events for Background 2. The large fractional uncertainty on Background 2 leaves open many possibilities of how to interpret this prediction. It is a frequent occurrence in HEP experiments to have at least one background component that has a large fractional uncertainty evaluated for its prediction. Such ill-constrained predictions have less of an effect on experimental results if they constitute a small amount of background, where small is in relation to the other background components or to the signal that is tested. The signal rate is left unspecified, and varies from one simulated dataset to another.

In each of the Challenge datasets, a rate was chosen for Background 1, Background 2, and the signal, based on the hypothesis under test. Truncated Gaussian distributions were sampled for the Background 1 and Background 2 rates. The seven signal hypothesis categories are listed in Table 2. A Poisson random number was chosen using the randomly chosen rates, and then marks were generated using the

Table 2: Problem 2 dataset categories – signal rates and how many repetitions of each were represented in the Challenge datasets. For the category marked with a "*", the participants were asked to compute their expected correct-discovery rates. The column headed n_{rep} lists how many simulated datasets were supplied for each of the categories.

Category #	Input Signal	n_{rep}
1	0.00	17600
2*	75.00	400
3	50.00	400
4	25.00	400
5	100.00	400
6	150.00	400
7	125.00	400

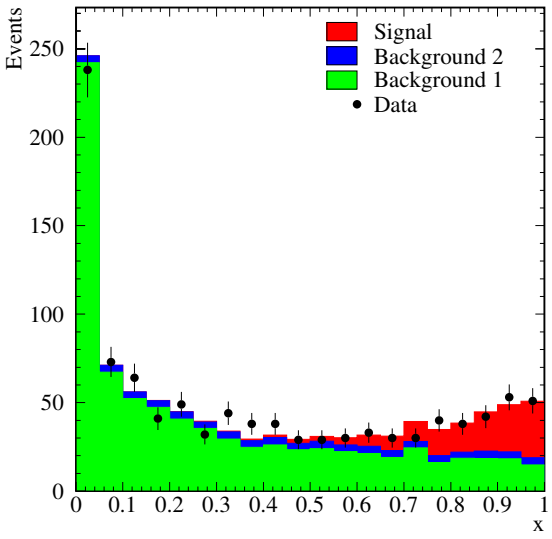

Fig. 2: An example distributions for Problem 2. The simulated data are displayed in a binned histogram, indicated by points with error bars drawn as $\sqrt{n_i}$ where n_i is count of simulated data events with marks that fall in that bin. The background and signal intensities are shown also with binned histograms formed from the Monte Carlo mark distributions for their respective samples.

prescriptions described above. The resulting lists of marks were then shuffled. The list of which simulated dataset was drawn from which signal test case was also shuffled.

Category 2 in Table 2 was chosen as the standard reference signal model for which the participants were to estimate their correct-discovery fractions.

Figure 2 shows an example distribution of marks for a simulated dataset with a prominent injected signal.

3 Submissions

Submitters were asked to describe briefly their methods, and their descriptions are reproduced below, alphabetized by the author's last name. There is some variety in the notation used.

3.1 Mark Allen

For Problem 1, Mark Allen provided a solution based on an unbinned maximum-likelihood fit, $\Delta \log \mathcal{L}$ as the test statistic for computing p values. In order to find the global maximum of the likelihood most often, several fits are performed with different starting conditions. The p values are computed by comparing a dataset's test statistic with a distribution of a large number of simulated background-only datasets. Since a signal can be found anywhere in the distribution on any of the simulated background-only datasets, the LEE is taken into account.

3.2 Stefano Andreon

For Problem 1, Stefano Andreon provided a solution based on a Bayesian computation with uniform priors on A and D, with a zero value for the prior for negative (unphysical) values, and an uniform prior, between 0 and 1, on E. Stefano computes $p(D = 0|\mathrm{data})$, up to a multiplicative factor, and selects simulated datasets for discovery claims if $p(D = 0|\mathrm{data}) < 3 \times 10^{-3}$ for the first solution, and $p(D = 0|\mathrm{data}) < 4 \times 10^{-3}$ for the second. The Type-I error rate is higher for the second set, but the power is also larger. An estimate of the power of the tests was not supplied.

3.3 Frederik Beaujean

For Problem 1, Frederik Beaujean and the Bayesian Analysis Toolkit (BAT) team provided a solution based on BAT's fast Poisson p value estimation, corrected for the number of degrees of freedom. The value of A that maximizes the posterior probability in the background-only case is used. If the p-value is less than 0.01, a Bayesian analysis is conducted, and a discovery is claimed if $P(B|\mathrm{Data}) < 0.001$. The LEE is taken into account by assuming a prior that favors the background model. A rather small fraction of the simulated datasets with injected signals had a discovery claim using this technique.

3.4 Matt Bellis and Doug Applegate

Matt and Doug's solution to Problem 2 involves a fit to each dataset using a nearest-neighbors algorithm to estimate the PDFs of the two background and one signal MC components, and a bootstrapping procedure to marginalize over the correlated uncertainties inherent in this approach. Matt and Doug use toy Monte Carlo studies to estimate the Type-I error rate and the power of this procedure.

The nearest-neighbor estimation of the PDF for each process is designed to evaluate the PDF at values of the mark x where there are data events, but not for arbitrary values of x. This reduces the problem from estimating a function to estimating a set of numbers $P^k(\vec{x}) = \{P^k(x_i)\}$ for the kth contributing process (one of the two backgrounds or the signal). For each data point x_i and channel k, Matt and Doug calculate a probability density by counting the number of nearby Monte Carlo samples N_s^k within a range r_s, then divide by r_s and the total number of Monte Carlo samples N_{tot}^k. The estimate of the PDF will have noise from the finite size of the Monte Carlo sample. In addition, the values $P(x_i)$ will be correlated since the same points are used for neighboring density estimations. To account for these effects, Matt and Doug produce bootstrap realizations of the Monte Carlo and calculate PDFs for each random draw. The ensemble of $\{P_j^k(\vec{x})\}$, where j indexes bootstraps, are random draws from the PDF of $P^k(\vec{x})$ that includes the uncertainty from both finite number and correlation.

Matt and Doug use the MINUIT minimizer [5] to find the best-fit fractions for the three processes by minimizing the negative log likelihood, which includes the PDF information above for each process, as well as Gaussian constraints on the rates from the problem specification. Matt and Doug elect to compute the ratio of the probabilities of each model at the best fit parameters, *i.e.* the delta log-likelihood. They calibrate the delta log-likelihood statistic by computing its distribution in zero-signal toy Monte Carlo simulations and use that to compute p-values and sensitivities.

3.5 Georgios Choudalakis

The BUMPHUNTER [6] is a hypothesis test sensitive to local excesses of data with respect to the null hypothesis. It is configurable, which means it can also be sensitive to deficits, and can optionally require agreement in the sidebands around the local discrepancies it evaluates. It is not assuming any specific shape for the potential discrepancy it is looking for, and by default it does not assume any specific width either, therefore it is highly model-independent. The trials factor associated with checking for discrepancies of various widths at various positions is taken into account using pseudo-experiments. The algorithm is fast enough to make the use of pseudo-experiments practical in most cases, without excluding the possibility of analytic approximations if needed. Understanding the BUMPHUNTER as a hypertest [6] allows for straight-forward generalizations, such as looking for discrepant tails in distributions (TAILHUNTER [6] [7]) or other features, and combining multiple datasets in a single hypothesis hypertest while accounting for the LEE.

3.6 Eilam Gross and Ofer Vitells

Eilam and Ofer provided a solution to Problem 1 based on a two-fit log likelihood ratio similar to those used by other participants. The LEE is addressed using a procedure described in [8]. It was found that the submitted p-value distribution extended well above 1.0, although this is not a problem for discovery. This method of addressing the LEE works well for small p-values which are required for discovery, but rather conservatively magnifies large p-values. Confidence intervals for D and E are computed using the likelihood ratio test $\Delta 2 \log \lambda = 1$, additionally setting the lower bound on the signal rate to be zero when $P(q_0 \leq q_0^{observed} | H_0) = 68\%$, where $q_0 = -2 \log \lambda(0)$ if the best-fit signal rate is positive, and zero otherwise, and

$$\lambda(N_s) = \frac{\mathcal{L}(N_s, \hat{\hat{N}}_b, \hat{\hat{M}})}{\mathcal{L}(\hat{N}_s, \hat{N}_b, \hat{M})},$$

where N_s and N_b are signal and background yields, and the hats indicate best-fit parameters. Eilam and Ofer provided a solution to Problem 2 using a likelihood ratio test statistic similar to that of Problem 1, except in this case the likelihood ratio is binned, and there is no LEE.

3.7 Tom Junk

For Problem 1, Tom provided a solution based on an unbinned profile likelihood test statistic. Two fits are done, both using MINUIT [5], one in the test hypothesis, and one for the null hypothesis. Simulated datasets were generated using the prior-predictive ensemble. The LEE is incorporated by testing all datasets in the same way, allowing a peak to be found anywhere in the ranges $0 < E < 1$ and $0 < D$. Tom reports the values of D and E returned by the MINUIT fit.

Tom provided a solution to Problem 2 using a binned likelihood technique. Aside from the binning, and the lack of a peak position parameter, the method used is very similar to the solution used for Problem 1. An additional feature is the limited sample size of the Monte Carlo used to predict backgrounds. This adds an extra nuisance parameter for each bin for each sample – signal, background 1, and background 2. Tom fluctuates all of the nuisance parameters in each of his simulated data samples used to characterize the test statistic. This differs from the prior used to generate the datasets in that the characterizing datasets are binned, and the priors in each bin are taken as Gaussian approximations to the distributions of the bin-by-bin parameters. A possibly better choice is to use a Gamma prior in each bin for the bin-by-bin uncertainties, which is the Bayesian result using the finite Monte Carlo and a uniform prior in the unknown true background and signal rates. This however biases the prediction upward in each bin. Tom fit the two background rates, but did not fit the separate bin-by-bin uncertainties, to get values of the $-2 \ln Q$ test statistic for the simulated datasets and the challenge datasets, where Q is the ratio of profile likelihoods under the test and null hypotheses.

For the signal rate intervals, Tom performed a Bayesian calculation, integrating the likelihood function times a uniform prior in the signal rate over the uncertain parameters (this time, the two background rates and the bin-by-bin uncertainties). The 68% credibility interval is computed as the shortest interval containing 68% of the integral of the posterior.

Since Tom had access to the correct answers for each simulated dataset, Tom's solutions are not eligible to "win" the competition.

3.8 Valentin Niess

Valentin's analyses of the two problems rely on frequentist hypothesis testing tools, but they differ with respect to the test statistic that is considered. The first algorithm counts the number of events within a subinterval of the possible range of marks $\Gamma = [0; 1]$ chosen in order to maximize the separation of signal from noise. The optimal half-width of the bracketing interval was found to be $\Delta = 1.4\sigma$ in this case, where $\sigma = 0.03$ is the signal width in E. The background contamination in the subsample is simultaneously estimated from the sidebands. The search for the best value of E is repeated over N_{bin} brackets overlapping over the range $[0; 1]$ by steps of $\delta = \sigma/2$. The LEE is taken into account by correcting the p value by an effective trial factor, given as: $N_{eff} = |\Gamma|/\sqrt{2\delta\Delta}$, where $|\Gamma|$ is the length of the interval Γ.

The second algorithm proceeds with the Kolmogorov-Smirnov (KS) statistic, parameterizing the signal and background cumulative distributions with power-law functions of the marks. The KS test statistic is minimized numerically over the uncertain values of the signal and background rates.

3.9 Wolfgang Rolke

Wolfgang's solution to both problems is based on the likelihood ratio test statistic

$$\lambda(\mathbf{x}) = 2\left(\max\{\log L(\theta|\mathbf{x}) : \theta\} - \max\{\log L(\theta|\mathbf{x}) : \theta \in \mathbf{\Theta}^0\}\right)$$

where $L(\theta|\mathbf{x})$ is the likelihood function.

According to standard theorems in statistics $\lambda(\mathbf{X})$ often has a χ^2 distribution in which the number of degrees of freedom is the difference between the number of free parameters under the test hypothesis and the number of free parameters under the null hypothesis. This turns out to be true for Problem 2 but not for Problem 1, in which case the null distribution can be found via simulation.

For Problem 1 the main difficulty is finding the maximum likelihood estimator (MLE) because the likelihood surface has a large number of local minima. To find the MLE, Wolfgang used a two-step procedure: first a fine grid search over values of the signal location E from -0.015 to 1 in steps of 0.005. At each value of E the corresponding value of signal size α that maximizes the log-likelihood is found. In a second step, Wolfgang starts at the best point found above and uses the Newton-Raphson method to find the overall MLE.

In Problem 2 the difficulty is in estimating the densities for the backgrounds and the signal from the available Monte Carlo data. Wolfgang explored the following solutions:

a) parametric fitting: for all three data sets Beta densities (with different parameters) yielded fits that passed a number of goodness-of-fit tests.

b) non-parametric: the densities are estimated using non-parametric kernel estimators.

c) semi-parametric: a combination of a) and b).

3.10 Stefan Schmitt

Stefan Schmitt analyzed both Challenge problems using the method of fractional event counting [9]. This method defines a test statistic $X = \sum N_i w_i$, where N_i is the observed number of events in bin i and

w_i is the fractional event weight. The weights w_i depend on the unknown nuisance parameters, namely the signal rate (problem 1 and 2) and the signal position (problem 1 only). The w_i are constructed from the expected signal and background contributions together with the size of the variations expected from systematic uncertainties. This calculation is done in a way which maximizes the sensitivity of X to the presence of a signal and at the same time minimizes the sensitivity of X to systematic variations [9].

Once the w_i are defined, the calculation of X is inexpensive in terms of computing power. Probabilities are thus calculated using Monte Carlo methods. In particular, the p-value is defined as the fraction of background Monte Carlo experiments which have a test statistic X larger than the one found for the experimental data. All nuisance parameters but the unknown signal properties are integrated over when generating the Monte Carlo experiments. For calculating the p-value and deciding on the presence or absence of a signal, the signal rate in the weight calculation is fixed to a value r_0. The rate r_0 is chosen such that the expected Type-II error for a signal with rate r_0 is approximately 50%.

For the case of Problem 1, the signal position E is not known. A scan is performed as a series of tests with variable E, where E is increased in the range 0 to 1 in steps finer than the signal resolution. The minimum p-value found in this scan is corrected for the LEE by repeating the scan on a sufficiently large number of independent Monte Carlo experiments.

3.11 Stanford Challenge Team

The SCT provided a solution to Problem 1 based on a log-likelihood ratio test statistic performing two fits to each dataset. The distribution of the test statistic is predicted using simulation. The LEE is handled by allowing any value of E to be fit in the simulated null hypothesis datasets used to calibrate the critical value. The parameters D and E were obtained using a maximum-likelihood fit. The SCT used the nonparametric bootstrap to estimate the variability of the results.

The SCT provided a solution to Problem 2 using a likelihood ratio test similar to that used in Problem 1, comparing a three-component fit to a two-component fit (three including the signal, and two backgrounds are fit in either hypothesis). The distributions of the marks for the two background components and the signal component were approximated with Beta distributions.

4 Performance Summary – Problem 1

Challenge participants generally did quite well discovering the signals that were hidden in the simulated datasets. Table 3 lists the measured Type-I error rates and the correct-discovery sensitivities for the submissions for Problem 1. Two calculations of the correct-discovery rates are listed. The "claimed" rate is estimated by the participant, and the "measured" rate is that obtained from the challenge datasets.

No one ignored the LEE – the Type-I error rates were nearly all under the desired 1%. Stefano Andreon's submission provided a Bayesian test statistic with two suggested cuts on it, both of which gave Type-I error rates in excess of the desired 1%. Adjustment of the cut can certainly produce the desired error rate, although at the price of fewer correct evidence outcomes. Stefano Andreon's submission has rather high discovery rates measured with the challenge datasets, and allowing tuning of the cut on the test statistic it is estimated that the performance of the Bayesian method is similar to that of the other methods.

There appears to be an upper limit on the correct-discovery fractions for each of the three standard signal hypotheses for which the participants were asked to evaluate their sensitivities, approximately 40%, 50%, and 20% (rounding up slightly), indicating that the choice of signal rates and positions was optimized well in order to make them measurable with fewer repetitions. No one participant had the highest claimed sensitivity for all three points and also a Type-I error rate under 1%. Figure 3 lists the fractions of signal-containing simulated datasets for which each participant claims evidence, for all 24 signal hypotheses (including the hypothesis of no signal). The author would like to thank Ofer Vitells for producing this figure and the following two.

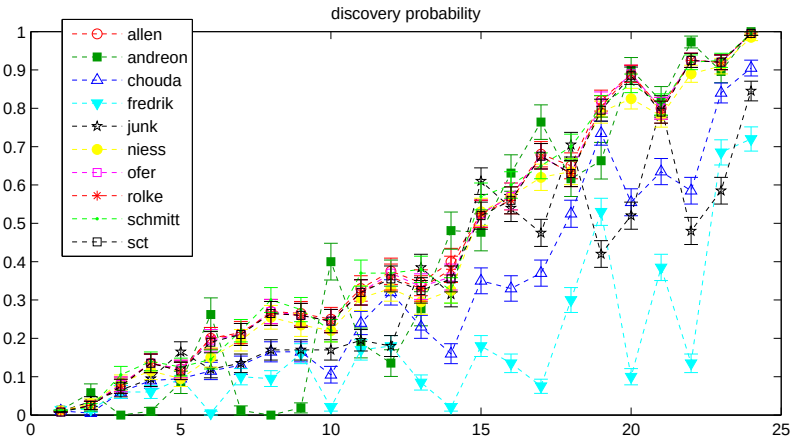

Fig. 3: Fractions of simulated datasets for which each participant claims evidence, for the 24 signal hypotheses of Problem 1 The categories are sorted on the horizontal scale according to the average correct-assignment fraction. The author would like to thank Ofer Vitells for preparing this figure.

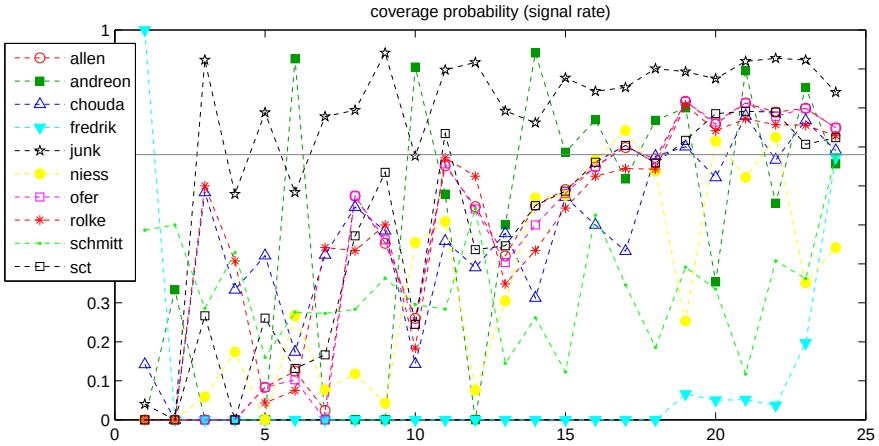

Fig. 4: Fractions of simulated datasets in which the 68% CL intervals quoted by the participants contain the true value of D, the signal strength parameter. The 24 signal categories are sorted in order of increasing average correct assignment probability. The author would like to thank Ofer Vitells for preparing this figure.

The performance for extracting the parameters D (which controls the signal strength) and E (which controls the signal position) were not as good. The summary note [4] provides listings for each participant for each signal model, the fraction of the datasets that claim evidence that also have intervals for D that contain the true value, and separately also for E. The average lengths of these intervals are also listed. Distributions of the fitted values of E, and the upper and lower edges of intervals of D are also shown in that note. Figure 4 in this article shows the fraction of the simulated datasets in which the signal rate parameter D is within the quoted intervals, which ideally should be 68% or greater. Figure 5 shows the fraction of the simulated datasets in which the signal position parameter E is within the quoted intervals, which also ideally should be 68% or greater. The true values of the signal rates and locations for each simulated dataset are provided on the web page [4] as an aid in investigating coverage issues with the intervals supplied.

Table 3: Listing of the claimed and measured correct-discovery rates for the three scenarios of Problem 1. Stefan Schmitt states that his unbinned sensitivities are rather similar to his binned sensitivities. The Bayesian technique proposed by Stefano Andreon did not have estimated correct-discovery rates provided.

Contributor	Type-I Error Rate	$D = 1010$,$E = 0.1$		$D = 137$,$E = 0.5$		$D = 18$,$E = 0.9$	
	Measured	Claimed	Measured	Claimed	Measured	Claimed	Measured
Tom Junk	0.0097 ± 0.0008	0.256	0.3150 ± 0.0328	0.543	0.6100 ± 0.0345	0.108	0.1350 ± 0.0242
Wolfgang Rolke	0.0103 ± 0.0008	0.356	0.3800 ± 0.0343	0.457	0.5250 ± 0.0353	0.184	0.2150 ± 0.0290
Stanford Challenge Team (SCT)	0.0077 ± 0.0007	0.3483	0.3550 ± 0.0338	0.4335	0.5200 ± 0.0353	0.0175	0.2100 ± 0.0288
Eilam Gross & Ofer Vitells	0.0082 ± 0.0007	0.35	0.3600 ± 0.0339	0.46	0.5250 ± 0.0353	0.19	0.2100 ± 0.0288
Valentin Niess	0.0111 ± 0.0008	0.34	0.3250 ± 0.0331	0.46	0.5300 ± 0.0353	0.17	0.1950 ± 0.0280
Georgios Choudalakis	0.0110 ± 0.0008	0.213	0.1600 ± 0.0259	0.290	0.3500 ± 0.0337	0.107	0.1300 ± 0.0238
Mark Allen	0.0106 ± 0.0008	0.385	0.4000 ± 0.0346	0.486	0.5250 ± 0.0353	0.187	0.2100 ± 0.0288
Frederik Beaujean (BAT)	0.0000 ± 0.0000		0.0000 ± 0.0000		0.0300 ± 0.0121		0.0050 ± 0.0050
Stefan Schmitt							
Unbinned	0.0112 ± 0.0009		0.4500 ± 0.0352		0.5450 ± 0.0352		0.1850 ± 0.0275
Binned	0.0110 ± 0.0008	0.37	0.3850 ± 0.0344	0.53	0.5450 ± 0.0352	0.17	0.2200 ± 0.0293
Stefano Andreon							
$p < 3 \times 10^{-3}$	0.0126 ± 0.0013		0.4811 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120
$p < 4 \times 10^{-3}$	0.0191 ± 0.0016		0.5189 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120

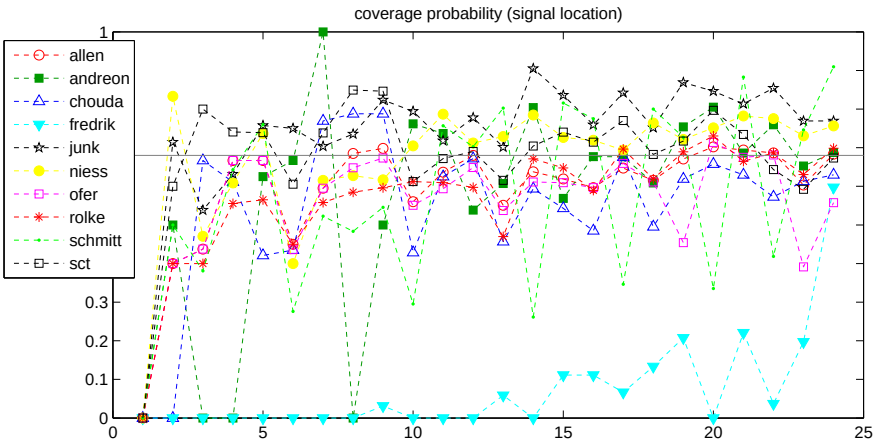

Fig. 5: Fractions of simulated datasets in which the 68% CL intervals quoted by the participants contain the true value of E, the signal location parameter. The 24 signal categories are sorted in order of increasing average correct assignment probability. The author would like to thank Ofer Vitells for preparing this figure.

5 Performance Summary – Problem 2

The solutions to Problem 2 had a broader spectrum of performance than for Problem 1. This is largely due to the ambiguity in specifying the model of the signal and background when only a Monte Carlo model is available. In practice, all an experimentalist has on hand for many signal and background predictions are finite Monte Carlo samples simulated to model these processes. If the size of the Monte Carlo samples is inadequate, larger samples can usually be generated. The data sample helps constrain systematic mismodeling features of the Monte Carlo simulation by comparing observed and expected rates and distributions for events which fail the main selection requirements but pass others designed to select events that can test the features of the Monte Carlo. For Challenge Problem 2, the average total number of background events is of order 1000, and so corresponding Monte Carlo model samples of size 5000 are in a typical ratio to actual data.

Some of the submissions parameterized the distributions of the marks for the signals and backgrounds using analytic functions, and others used binned likelihoods. It was found that the methods that parameterized the shapes as beta functions tended to undercover on the Type-I error rate. The beta function parameterization likely mis-predicts the density of the marks near zero. Participants that ran into this issue re-performed their analysis of the simulated datasets using their method and the true spectrum once the results were distributed, and achieved an error rate of 1%. Coverages for the signal rate fits are provided on the problem web page [4].

6 Summary

The solutions to the Banff Challenge 2 problems provided by the participants span a range of different approaches. Most of the hypothesis tests are based on ratios of profile likelihoods, with Monte Carlo simulation of the distribution of the test statistic. Minor variations between submissions arise from the choice of binning or unbinned fits, and the strategy used to find a global minimum among many local minima in the first problem, and in the parameterization and handling of the distributions of the marks in the second problem. Alternate approaches involved counting events inside signal windows while fitting backgrounds in the sidebands, counting fractional events, and using the Kolmogorov-Smirnov test statistic, and Bayesian methods. Bayesian methods do not naturally focus on error rates, which are frequentist concepts, making the problem setup somewhat clumsy for Bayesian analysis.

The Look-Elsewhere Effect is an issue in Problem 1 (but not in Problem 2) since the presence of

Table 4: Listing of the Type-I error rates, and the claimed and measured correct-discovery rates for the signal scenario Problem 2 for which the participants were asked to estimate their discovery power. Stefan Schmitt states that the power of his 50-bin test is similar to that of his 25-bin test.

Contributor	Type-I Error Rate	Signal = 75 Events	
	Measured	Claimed	Measured
Tom Junk	0.0068 ± 0.0006	0.865	0.870 ± 0.017
Wolfgang Rolke	0.0256 ± 0.0012	0.88	0.8500 ± 0.018
Stanford Challenge Team	0.0389 ± 0.0015	0.84	0.9100 ± 0.0143
Eilam Gross & Ofer Vitells	0.0107 ± 0.0008	0.815	0.7725 ± 0.0210
Valentin Niess	0.0085 ± 0.0007	0.761 ± 0.001	0.7125 ± 0.0226
Stefan Schmitt			
25 Bins	0.0047 ± 0.0005	0.85	0.8200 ± 0.0192
50 Bins	0.0047 ± 0.0005		0.8250 ± 0.0190
Doug Applegate & Matt Bellis	0.0168 ± 0.0010	0.95	0.8950 ± 0.0153

a signal introduces an additional parameter – the location of the peak E in the test hypothesis which is not present in the null hypothesis. All participants handled this effect rather well – there are no signs of noticeable undercoverage in the Type-I error rate measurements. One of the methods of accounting for the LEE had the effect of producing p values in excess of unity however.

A typical HEP experiment uses a flip-flopping approach to decide when to quote a two-sided interval and when to quote a one-sided upper limit. The part of the Challenge specification asking for two-sided intervals when evidence was claimed and otherwise not did not allow a unified approach. This request biased the intervals on the rate parameter upwards, most noticeably in the simulated datasets drawn from the null hypothesis. Quoting a two-sided interval for the production rate of a new particle for which evidence is not claimed can be misconstrued by the broader community, even though doing so would help the coverage properties of the methods. Nonetheless, most of the methods provided solutions that undercovered for the signal rate and location parameters for Problem 1.

It is in Problem 2 that significant undercoverage in the main result, quoting evidence or not, meaning a higher-than-expected Type-I error rate, was seen in several submissions. Participants both underestimated their Type-I error rates and overestimated their discovery power. Because the distributions of the marks were not given to the participants, instead relying on simulated Monte Carlo samples of them, participants either binned the data or calculated unbinned likelihoods using parameterizations that appear to fit the distributions of the marks in the simulated Monte Carlo samples. If these parameterizations do not match the true distribution (and they are guesses since the true distribution is hidden), they could, and did, result in poor estimates of the Type-I and Type-II error rates. It could also be that the *a priori* uncertainty of 100% on the rate of Background 2 causes ambiguities to arise in the approach to follow that is reflected measurably in the results, particularly since Background 2 looks more like the signal than Background 1 does.

In general, the methods did very well – an impressive array of approaches, conscientiously applied, gave similar performances and for the most part met the specifications set forth in the Challenge. The Challenge's goal of giving practice on a realistic set of problems is well met. There are many different possible metrics for success on the Challenge problems, just as there are in a real HEP experiment, and no participant's solution came out on top in every possible metric. In a real HEP experiment, the statistical methods used for discovery and exclusion must be approved by the collaboration and reviewed by journal editors and referees. This Challenge provides useful practice in developing, applying, and characterizing

techniques which can be used to test for new phenomena.

Acknowledgements

The author would like to thank the Banff Challenge 2 team, Louis Lyons, Richard Lockhart, Jim Linnemann, and Wade Fisher. This work was performed at Fermi National Accelerator Laboratory, operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy.

References

[1] J. Heinrich, CERN Report CERN-2008-001, p. 125 (2008). http://cdsweb.cern.ch/record/1021125?ln=en

[2] L. Lyons, Annals of Applied Statistics **2**, 887 (2008);
L. Demortier, Proceedings of PHYSTAT-LHC 2007, CERN-2008-001.

[3] T. Junk, http://www-cdf.fnal.gov/~trj/

[4] T. Junk, http://www-cdf.fnal.gov/~trj/bc2sub/bc2sub.html

[5] F. James, CERN Program Library Long Writeup D506, Version 94.1 (1998). http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html.

[6] G. Choudalakis, [arXiv:1101.0390 [physics.data-an]].

[7] ATLAS Collaboration, Phys. Rev. Lett. **105**, 161801 (2010). [arXiv:1008.2461 [hep-ex]].

[8] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", Eur. Phys. J. C **70**, 525 (2010).

[9] P. Bock, JHEP **0701** (2007) 080 [arXiv:hep-ex/0405072].