

# A Submission for the Banff Challenge 2a Problems

Thomas Junk  
*Fermi National Accelerator Laboratory*

December 29, 2010

## Abstract

This note summarizes the techniques used to solve the Banff Challenge 2a problems, which test the procedures for finding evidence for and discovering new signals amid large, *a priori* uncertain backgrounds. The discovery decisions are based on *p*-values calculated within the prior-predictive ensemble, that is, integrating over the *a priori* distributions of the nuisance parameters. The effects of the uncertainties on the sensitivity of the search is reduced by fitting for the uncertain parameters. Intervals are calculated for the signal rates in Problems 1 and 2, and intervals for the peak position are calculated for Problem 1.

## 1 Introduction

The problems are specified in a separate note, available at

<http://www-cdf.fnal.gov/~trj>.

This document summarizes an entry to the challenge problems, and describes its performance on simulated datasets – not the challenge datasets but additional ones used to characterize the methods.

## 2 Challenge Problem #1

For Challenge Problem #1, The solution provided is based on an unbinned profile likelihood test statistic.

$$L = e^{-(A-10000)/2 \cdot 1000^2} \prod_{i=1}^{n_{\text{events}}} \left[ A e^{-cx_i} + D e^{-(x-E)^2/2\sigma^2} \right] \quad (1)$$

Two fits are then done to each simulated dataset. The first one maximizes  $L$  with respect to  $A$ , setting  $D = 0$ , and is called the “background-only fit”. The second one maximizes  $L$  letting  $A$ ,  $D$ , and  $E$  float simultaneously, and is called

the “signal-plus-background fit”. The first fit does not depend on the values of  $x_i$  on the data events, but simply is a count.  $A$  is then a weighted average of 10000 and  $10n_{\text{events}}$ , where the weights are given by the quoted uncertainties of  $\pm 1000$  on  $A$  from the external constraint, and  $\pm 10\sqrt{n_{\text{events}}}$  for the data measurement. This latter fit is performed with MINUIT. The dimensionality of the parameter space is reduced by fitting for the fraction of signal events  $f_{\text{sig}}$  and  $E$ , where

$$A = C \times n_{\text{events}} / (1 + f_{\text{sig}}), \quad (2)$$

taking advantage of the fact that the constraint from the total number of events is much stronger than the prior constraint. To be fully general, the three parameter fit should be employed but it is less numerically stable and takes more computer time.

The test statistic is

$$\Delta \log L = L_{\text{background}} - L_{\text{signal}} \quad (3)$$

Nine million background-only pseudoexperiments were drawn using the prior-predictive sample space. Each sample was simulated first by drawing  $A$  from a normal distribution of mean 10000 and width 1000, and from that, a Poisson total number of events was randomly chosen, and with that total, an exponentially distributed set of  $x_i$  were randomly chosen, assuming  $C = 10$ . The distribution of  $\Delta \log L$  was computed for the background events and it is shown as the black solid histogram in Figure 1. This distribution was used to compute the  $p$ -value of each of the data outcomes. The LEE is incorporated implicitly here since each simulated background outcome is fit allowing the peak position to be anywhere in the signal hypothesis. In the signal-plus-background fit, MINUIT’s “seek” function was first called, and then “minimize”.

Tom found that 1% of background-only simulated datasets (not the challenge datasets but his own thrown from the prior-predictive distribution) have  $\Delta \log L < -0.508203$ , which is shown as the blue dotted line in Figure 1. Signal-plus-background pseudoexperiments, three million for each of the signal test scenarios, were drawn in a similar way as the background-only pseudoexperiments, first drawing  $A$  from a random distribution, and then drawing random Poisson data counts from the signal and smeared background predictions, and then drawing random data  $x_i$ ’s from the signal-plus-background distributions. The distributions of  $\Delta \log L$  for the signal-plus-background simulated datasets are also shown in Figure 1. The power is computed as the number of signal-plus-background outcomes that correctly result in evidence being quoted, with a  $p$ -value less than 0.01. For the first scenario, with  $E = 0.1$ ,  $D = 1010$ , Tom finds a correct-discovery rate of 0.256, or a Type-II error rate of 0.744. For the second scenario, with  $E = 0.5$ ,  $D = 137$ , the correct discovery rate is 0.543, corresponding to a Type-II error rate of 0.457, and for the third scenario, with  $E = 0.9$ ,  $D = 18$ , the correct discovery rate is 0.108, corresponding to a Type-II error rate of 0.892.

To measure the event rate and peak position, the best-fit values from MINUIT and MINUIT’s uncertainties were used. Figure 2 shows the distribution of fitted

Table 1: Correct evidence fractions for the three working points for Tom Junk’s solution to Problem 1

$D$	$E$	Correct Evidence Fraction
1010	0.1	0.256
137	0.5	0.543
18	0.9	0.108

peak positions for signal-plus-background simulated datasets (not the challenge datasets but the ones used to calibrate the  $p$ -values), separately for the three signal scenarios proposed in the original problem statement. Only those outcomes giving a  $p$ -value less than 0.01 are shown.

The correct-evidence fraction at the 1% Type-I error rate is evaluated by finding the critical value of  $\Delta \log L$  which gives a 1% error fraction in background-only repetitions (each one with an independently chosen value of the one nuisance parameter), and then finding the fraction of signal-plus-background repetitions have  $\Delta \log L$  less than or equal to this critical value. The critical value is found to be -5.08203, and the performances are listed in Table 1.

### 3 Challenge Problem #2

A solution to problem 2 is provided by binning the data marks in 20 bins from 0 to 1, forming signal, two background, and data histograms. The signal and background histograms were normalized to the rates specified in the problem. Bin-by-bin priors were taken from  $\sqrt{n}$  variations in the Monte Carlo predictions, and multiplicative truncated Gaussian priors were used for the overall normalization. For each simulated dataset, the test statistic

$$-2 \ln Q = -2 \frac{L(\text{data}|s + b, \hat{\nu})}{L(\text{data}|b, \hat{\nu})} \quad (4)$$

is computed, where  $\nu$  indicates the set of nuisance parameters, and the hats on  $\nu$  indicate two fits, one assuming that a signal is present at the rate specified in the sensitivity calculation specification – 75 events of signal expected (one hat) and the other fit assumes that the signal is absent. For computational speed, the bin-by-bin nuisance parameters are not fit for. The nuisance parameters are therefore the rates of backgrounds 1 and 2, and are considered independent.  $L$  is a product of Poisson probabilities over the 20 bins in the histogram, augmented with Gaussian constraints on the two nuisance parameters.

The distributions of  $-2 \ln Q$  for signal-plus-background and background-only simulated datasets – not ones supplied in the problem definition, but additional ones generated randomly to characterize the method – are shown in Figure 3. The procedure for generating the datasets for filling Figure 3 is to randomly choose values of the two background normalizations from the priors given in the problem statement, fluctuating the predictions within each bin according to

the bin-by-bin priors, and generating Poisson pseudodata from the fluctuated predictions. This is called the “prior predictive” ensemble.

The correct-evidence fraction is computed in the same way as it is for Problem 1. The critical value of  $-2 \ln Q$  is -4.3095, and the correct-evidence fraction at the stated working point (75 events of signal expected) is 86.5%.

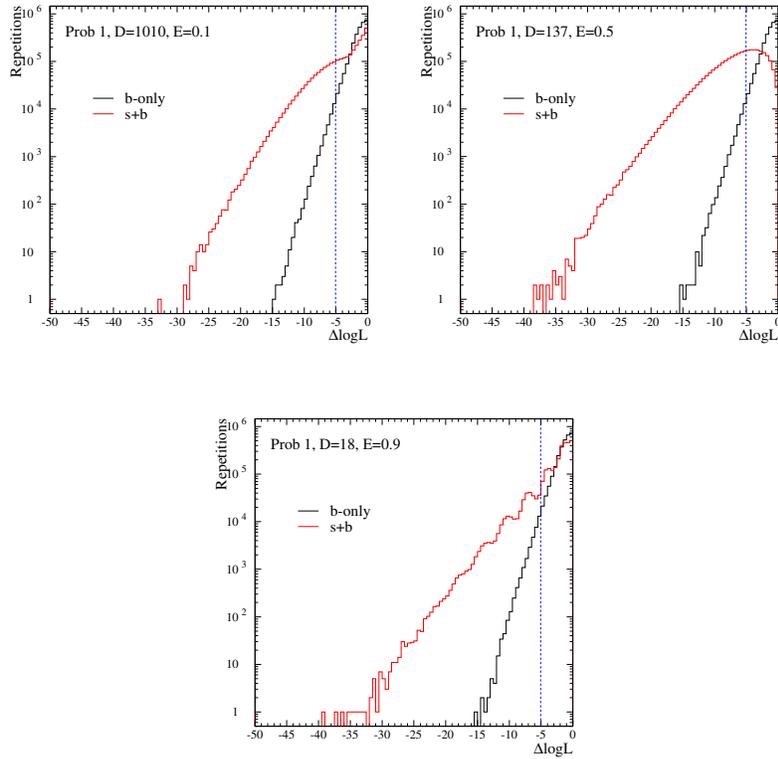


Figure 1: Distributions of  $\Delta \log L$  for Tom Junk's solution to Problem #1, for the three signal scenarios. The solid black histograms show the distribution in background-only simulated outcomes (Tom's own, not the challenge datasets), and the solid red histograms show the distributions for the three signal scenarios. The dashed blue line shows the critical value of  $\Delta \log L$  for which the Type-I error rate is 0.01. The third set shows the effect of the discreteness of individual data events.

## Prob 1, $\Delta \log L < -5.08293$

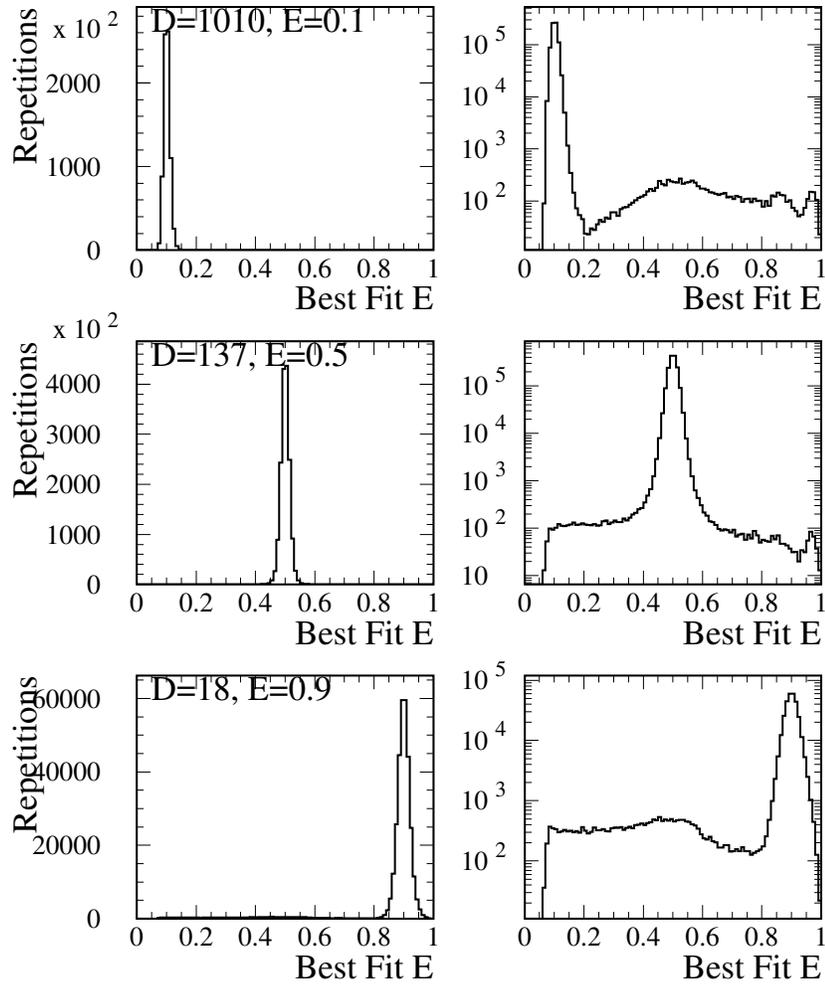


Figure 2: Best-fit values of  $E$  in a large sample of pseudo-datasets generated by Tom Junk to compute the power of the search for the three proposed signal scenarios. The left-hand plots have linear vertical scales, and the right-hand plots have logarithmic vertical scales.

### Problem 2, $s=75$ Events

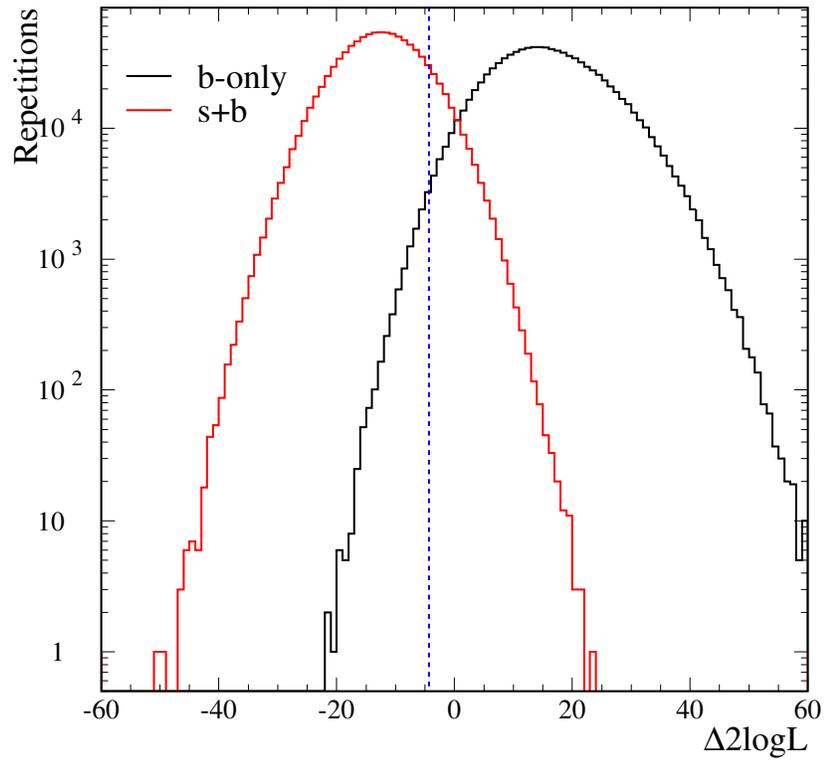


Figure 3: Distributions of  $-2 \log Q$  for Tom Junk's solution to Problem #2, for the one signal scenarios. The solid black histogram shows the distribution in background-only simulated outcomes (Tom's own, not the challenge datasets), and the solid red histogram shows the distribution for the signal scenarios. The dashed blue line shows the critical value of  $\Delta \log L$  for which the Type-I error rate is 0.01.