Banff Discovery Challenge Instructions

April 8, 2010

Contents

1 Introduction

The Banff discovery challenge is designed as a series of realistic data analysis problems similar to what may be experienced at the LHC or in analogous physics experiments. Some effort has been taken to provide each challenge problem in a format close to what is available to the experimentalists involved in the research. The challenge is a basic hypothesis-testing problem and the goal is to calculate the significance of a potential signal source in an observed measurement. Information on shapes and absolute normalizations for each challenge problem are provided and should be included in calculations of signal significance. This document uses references and nomenclature given within the general description of an experimental measurement provided by Richard Lockhart. This description linked together with the document you are now reading (entitled: Banff Challenge 2, Richard's Interpretation).

2 Problems for Banff Discovery Challenge

Each problem consists of a set of data that represent three distinct samples:

1. an observation (experimental result)
2. a background prediction
3. a signal prediction

Each data sample consists of "events" which are described by a variable corresponding to a measured quantity (observable). In order to provide a descriptive model for the background and signal predictions, these data samples are provided with high-statistics (500k events per sample). The number of observed data events corresponds to the experimentally recorded number of events. The predictive samples for the background and signal represent the experimental technique used to determine the intensity ($\lambda(x)$) for distinct predictions of event sources. The probability densities for these processes ($f(x)$) are derived from the intensity. For example, the probability density for the background process is defined as

$$f_b(x) = \frac{\lambda_b(x)}{\int \lambda_b(u)du} \tag{1}$$

The predictions for the signal and background processes require a normalization procedure to obtain the total expected number of predicted events. The total rate for the signal and background will be determined by three parameters: selection efficiency/acceptance (ϵ), luminosity ($\mathcal{L}$) and the cross section (β):

$$N_{\text{events}} = \beta \times \mathcal{L} \times \epsilon \times \int f(x)\, dx \tag{2}$$

where $\int f(x)\, dx = 1$ by definition. The observed data require no normalization procedure and the number of observed events is simply the sum of the intensity (λ_d).

There are five potential nuisance parameters that may be assigned using a Gaussian prior. For simplicity, the first three each will be defined with a standard deviation at 10% of the nominal value for the following:

1. Uncertainty on the background cross section
2. Uncorrelated uncertainties on the background and signal efficiencies
3. Correlated uncertainty on the luminosity for signal and background

The final two nuisance parameters are associated with uncertainty on the signal and background probability densities (f_s and f_b). These uncertainties are described via alternative background and signal samples which designate a change in f_s and f_b. These alternative samples exist for both ± 1 standard deviation measurements (i.e., describing a 68% CL region for the nuisance parameter). Thus, each problem will consist of:

- 3 data samples as above, defining f_s, f_b, and the observation.
- 4 data samples defining alternative f_s and f_b (± 1 standard deviation for each f_s and f_b)
- 2 cross sections, one for signal and one for background
- 2 efficiencies, one for signal and one for background
- 1 luminosity, common for signal and background
- Background cross section uncertainty uncertainty standard deviation (Gaussian standard deviation = 10%)

- Background efficiency uncertainty standard deviation (Gaussian standard deviation = 10%)
- Signal efficiency uncertainty standard deviation (Gaussian standard deviation = 10%)
- Luminosity uncertainty standard deviation (Gaussian standard deviation = 10%)

2.1 Notes on sample normalizations

Cross section values are provided in the unit of barns (b). Specifically, for this problem we will use the unit femtobarn (femto = 10^{-15}). These cross section values represent the probability for a physical process to occur per interaction, normalized to a fixed number of possible interactions. Luminosity values are provided in inverse femtobarns (b^{-1}) and represent the number of possible interactions as recorded by the experimental apparatus. Efficiencies are provided as an absolute fraction between 0 and 1, and represent both the fraction of events within the experimental acceptance and also the fraction of events within the kinematic selection utilized for the data analysis. Cross sections and efficiencies are specific to each process (i.e., background and signal processes each have unique values for both). The value of the luminosity is specific to the experiment and is common for both the signal and background processes.

More information on luminosity and cross section units can be found here:

- http://en.wikipedia.org/wiki/Barn_(unit)
- http://en.wikipedia.org/wiki/Luminosity

The data samples used to describe the intensities λ_s and λ_b each consist of 500k events. Following the normalization procedure in Eqn. 2 and example values $\mathcal{L}$=1000, $\beta = 2$, and $\epsilon = 0.5$, we find a value of 1000 expected events for the example process. As noted above, the observed data require no normalization procedure.

2.2 Description of Challenge Problems

Problem 1

In this problem, the data represent a standard multivariate event classification technique with events assigned a test statistic between 0 and 1: background is classified to values near 0 and signal is classified to values near 1. The specific parameters for problem 1 are:

- $\mathcal{L} = 100$ inverse femtobarns
- $\beta_{\text{signal}} = 4.2$ femtobarns
- $\beta_{\text{background}} = 2000$ femtobarns
- $\epsilon_{\text{signal}} = 0.05$
- $\epsilon_{\text{background}} = 0.50$

After normalization, one should expect $N_{\text{background}} = 10,000$ events and $N_{\text{signal}} = 210$ events. The observed number of events is 9815. A histogrammed distribution of the signal, background, and observed events for problem 1 is shown in Fig. 1.

Problem 2

In the second problem, the data represent a classic "mass bump" search, with a exponentially falling background and a peaked signal. The specific parameters for problem 2 are:

- $\mathcal{L} = 100$ inverse femtobarns
- $\beta_{\text{signal}} = 4.0$ femtobarns
- $\beta_{\text{background}} = 2000$ femtobarns
- $\epsilon_{\text{signal}} = 0.05$
- $\epsilon_{\text{background}} = 0.50$

After normalization, one should expect $N_{\text{background}} = 10,000$ events and $N_{\text{signal}} = 200$ events. The observed number of events is 9843. A histogrammed distribution of the signal, background, and observed events for problem 2 is shown in Fig. 2.

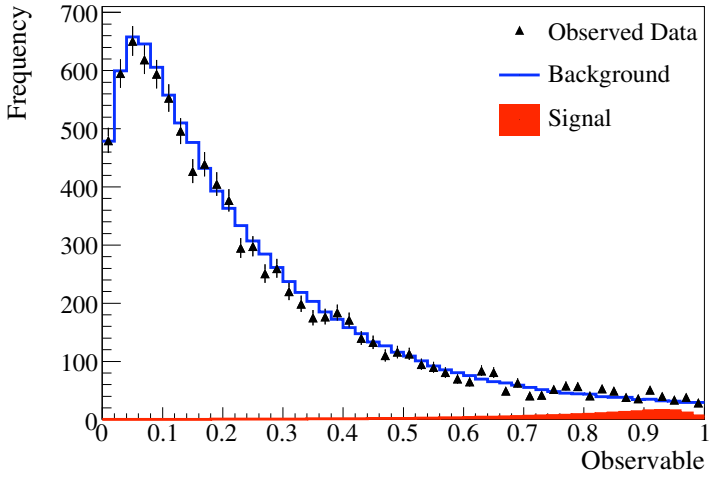

Figure 1: Histogrammed data for challenge problem 1 after applying normalizations. Shown in the figure are distributions for the observed data, the background prediction, and the signal prediction.

Problem 3

For the third problem, the data represent a search for a rare process with irreducible backgrounds in which the background and signal are relatively indistinguishable in the test statistic provided ($f_s \simeq f_b$). The specific parameters for problem 3 are:

- $\mathcal{L} = 10$ inverse femtobarns
- $\beta_{\text{signal}} = 18$ femtobarns
- $\beta_{\text{background}} = 800$ femtobarns
- $\epsilon_{\text{signal}} = 0.40$
- $\epsilon_{\text{background}} = 0.01$

After normalization, one should expect $N_{\text{background}} = 72$ events and $N_{\text{signal}} = 80$ events. The observed number of events is 134. A histogrammed distribution of the signal, background, and observed events for problem 3 is shown in Fig. 3.

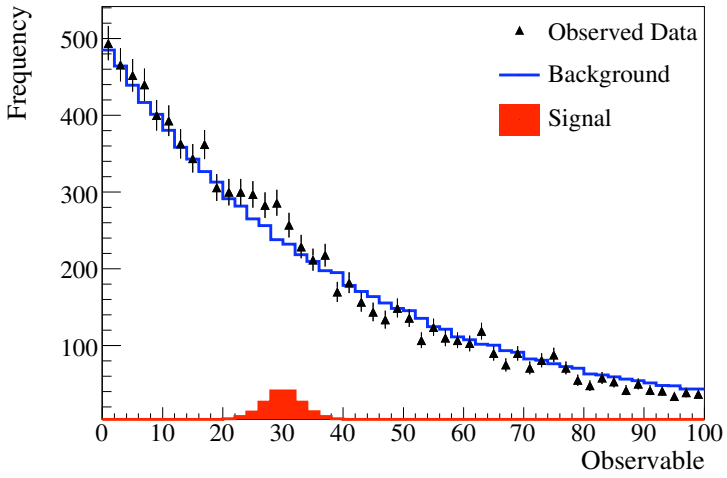

Figure 2: Histogrammed data for challenge problem 2 after applying normalizations. Shown in the figure are distributions for the observed data, the background prediction, and the signal prediction.

3 Banff Discovery Challenge Program

As the breadth of the challenge problems described above is sufficiently large, the challenge is divided into three distinct aspects in order to provide manageable thresholds of effort. These different aspects of the challenge should be considered to be an open-ended, ongoing project that will be discussed at the Banff conference. Preliminary results from the challenge can be presented and discussed, with a plan for further investigations emerging from these discussions.

Aspect 1: No Nuisance Parameters

The first aspect is the interpretation of each challenge problem in the absence of nuisance parameters.

Aspect 2: Introduce Nuisance Parameters

Repeat aspect 1 with nuisance parameters as described above.

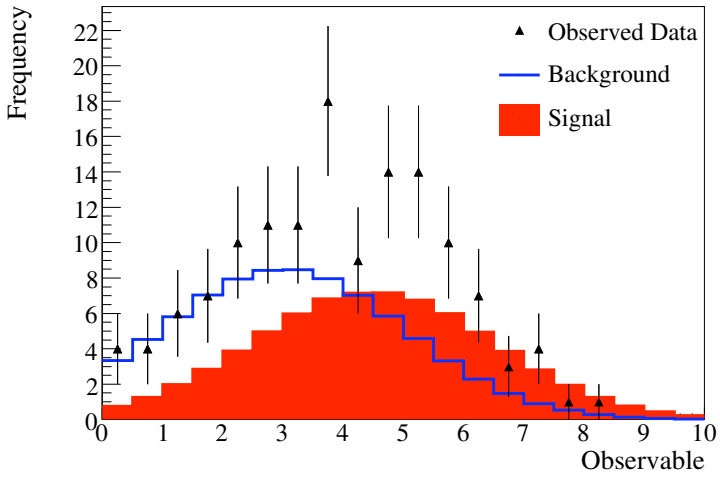

Figure 3: Histogrammed data for challenge problem 3 after applying normalizations. Shown in the figure are distributions for the observed data, the background prediction, and the signal prediction.

Aspect 3: Combine "Channels"

Combine the data in challenge problems 2 and 3 to find a joint calculation of the signal significance. For simplicity, assume all nuisance parameters are uncorrelated. The signals are to be assumed to arise from the same source. The two problems can thus be viewed as two independent results from two different experiments.

4 Files Provided

The data for the challenge problems are provided in two formats. First, the data are available in text format as sequential lists of events. These are found in the file banffChallengeText.tgz. Extract using the command "tar -xzf banffChallengeText.tgz".

The data are also available in ROOT format (http://root.cern.ch). There is a file for each challenge problem and each file contains histogrammed events as well as a TTree containing the individual events.

5 Questions and Requests for Information

Any questions or requests for information may be directed to Richard Lockhart (email: lockhart@stat.sfu.ca), Louis Lyons (email: l.lyons1@physics.ox.ac.uk), and Jim Linnemann (email: Linnemann@pa.msu.edu).