BANFF CHALLENGE 2

RICHARD'S INTERPRETATION

We imagine turning on an experiment and letting it run. As it runs events are generated. Each time an event is generated a set of associated measurements is collected. If we think of the set of possible measurement values for an individual event as a *mark space* then we have a marked Poisson process.

The number of events is typically huge. The events are preprocessed in software which has the effect of retaining only those events in some tiny part of the mark space; the preprocessing cuts (see the "Jargon" list at the end of these notes) the number of events to the range of say 10^1 to 10^3. We are going to model the retained events as follows. Calendar time is ignored[1]. There are N retained events; N has a Poisson distribution. The marks of the retained events are summarized into some rather simple numerical measurement denoted X; the summarization process is itself quite complex and typically involves outputs from some machine learning algorithm trained on very large Monte Carlo data sets.

We therefore have, in the end, data of the form $X_1, \ldots, X_N$ which take values in some set which can often be thought of as just the unit interval. The Xs are the points of a Poisson process on this set. I will denote the intensity of this process by $\lambda(x)$.

Our goal is the detection of some new phenomenon – the Higgs particle, say. We think of each X_i as having been produced either by well known phenomena, called background, or this new phenomenon. If the new phenomenon does not exist then the intensity is just $\lambda(\cdot) = \lambda_b(\cdot)$. If the new phenomenon exists then $\lambda(\cdot) = \lambda_b(\cdot) + \lambda_s(\cdot)$. The goal is to use the data N and $X_1, \ldots, X_N$ to decide between these two possibilities. It is likely that we will want to treat the background only model as a null hypothesis and require, in the ancient Neyman-Pearson way, strong evidence against the background only model before claiming discovery[2]. In summary the null hypothesis is $\lambda_s(\cdot) \equiv 0$.

We can attack the problem at various levels of realism. I make a list below with the simpler ones generally earlier in the list, I hope. Potentially important wrinkles are relegated to footnotes. The Banff Challenge 2 will be to compare methods for solving these problems

[1] In fact the rate at which events are produced is not really constant in time. It is proportional to "beam luminosity" which varies with time and integrated luminosity would replace calendar time. Moreover the rate at which interesting events are produced and the rate at which bad background events are produced both change with luminosity. These effects are often ignored.

[2] These notes focus on hypothesis testing. If we conclude that the signal is not 0 (and physicists would conventionally look for a one sided P-value $p < 2.85 \times 10^{-7}$ — 5 standard errors for a normally distributed test statistic) then there will be lots of interest in the values of the parameters; the original Banff challenge concerned confidence intervals for $\mu_s = \int \lambda_s(x)dx$. Physicists are very much interested in the coverage probabilities after such a preliminary test and in adjusting intervals to take account of the preliminary test.

on the basis of some artificial data sets. The various functions described in the Frameworks below are not know analytically, but are provided as (training) samples — outputs of large Monte Carlo runs. The Banff Challenge 2 problem will be discussed further at the workshop; if there is sufficient interest, methods may be compared quantitatively on more realistic versions of the problem. The problem is a stylized version of a typical real problem, as encountered, for example, at the LHC in the search for elusive Higgs Boson, but in many other contexts as well.

Framework 1: We begin by imagining that $\lambda_b(\cdot)$ is a known non random intensity[3]. In this case under the background model N has a Poisson distribution with mean $\mu_b = \int \lambda_b(x)dx$. Given N the $X_1, \ldots, X_N$ are an independent and identically distributed sample from the density

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

We might approach this then by testing both the hypothesis that N has mean μ_b and that the conditional density f of the X_i is f_b. The former problem is a standard one sided Poisson mean problem[4] and the latter is a goodness of fit problem since so far no clear form of the alternative density is proposed. The problem is to provide a single sensible summary P-value for testing this joint null; the P-value should be well calibrated (have as close as possible to a uniform distribution under the null or as close as possible to a known distribution under the null) and highly sensitive (that is have high power).

Framework 2: In Framework 1 we did not specify the signal intensity. That led to a classical goodness-of-fit problem. In fact there are often theoretical calculations of what the signal should look like if it exists. That is, under the alternative to the all background

[3] In fact it is subject to uncertainty of the "systematic errors" variety; see later Frameworks.

[4] It is standard and simple and there are therefore *many* suggestions for the solution and many suggestions for confidence intervals or limits for λ_μ. See Banff Challenge 1.

model we have the following specification:

$$\begin{aligned} N &\sim \text{Poisson}(\mu) \\ X_1, \ldots, X_N | N &\sim \text{iid with density } f \\ \mu &= \mu_b + \mu_s \\ \mu_b &= \int \lambda_b(x) dx \\ \mu_s &= \int \lambda_s(x) dx \\ f_b(x) &= \frac{\lambda_b(x)}{\mu_b} \\ f_s(x) &= \frac{\lambda_s(x)}{\mu_s} \\ f(x) &= \frac{\mu_b f_b(x) + \mu_s f_s(x)}{\mu} \end{aligned}$$

With this model and completely known intensities λ_s and λ_b we would just use the Neyman Pearson test if the intensities were computable and if we could compute the distribution of the likelihood ratio. This computation would automatically take care of the question of how to combine a test based on N with a test examining the conditional distribution of the X_i given N.

Framework 3: We can extend framework 2 by acknowledging that the background intensity λ_b is not exactly known[5]. Its value depends on measured features of the experimental set up and on computed values of various physical quantities. Both the measurements and the computations have uncertainties; giant Monte Carlo studies play an important role in the computations. From a statistician's perspective λ_b depends on many parameters. These parameters are measured with error and it is not easy and likely not possible to propagate these errors to modelled uncertainty in λ_b.

Typically μ_b might be regarded as having some sort of prior mean $\hat{\mu}_b$ and an uncertainty captured by some "systematic error variance" σ_b^2.[6] Usually only those two numbers are available, and the analyst is left to decide what information they represent about a putative prior model for $lambda_b$, though it will typically be agreed that the true $mu_b > 0$.

The problem cannot be approached as goodness of fit test for a parametric model with completely unknown parameters because the systematic errors are much more well characterized than could be estimated just from $X_1, \ldots, X_N$ (in spite of my remarks above about the vague information). We cannot hope to estimate parameters in λ_b just from these data and then test fit in the statistician's usual way. At the same time the systematic error is

[5] That is, in this framework the signal intensity λ_s is known. This is not realistic. More realistic is to suppose that f_s, the "shape" of λ_s, is known or fairly well known but that μ_s is not at all well known.

[6] Jim tells me "often has only vague information on $\hat{\mu}_b$ and $\hat{\sigma}_b^2$" but I am not sure what a Bayesian would make of this sort of uncertainty unless it means hierarchical priors are called for.

not so small that λ_b can be taken to be fully known. We probably need some prior model for λ_b which summarizes the uncertainties into a distribution on a random function and tests the joint hypothesis:

$$\begin{aligned} \lambda_b(\cdot) &\sim \text{Some random function / stochastic process model} \\ X_1, \ldots, X_n | \lambda_b &\sim \text{Poisson process with intensity } \lambda_b. \end{aligned}$$

Framework 4: If we add a parameter, say m, to the function λ_s and go back to Framework 2 we seem to have again a likelihood ratio problem. The null hypothesis is still "simple" in the language of Neyman and Pearson but the alternative is "composite". The parameters here are part of what we are trying to find out about. While we have some prior expectations about λ_s (which enable us to cut the data set back from $N =$ huge to $N = 10^2$ish) we want to use the measured events to pin down details. I am unclear about the extent to which a prior on λ_s would be tolerated.

Framework 5: We really want to add a fully or partially specified λ_s to Framework 3 leaving us with uncertainty both on the null and on the alternative about the intensity of the Poisson process. Statistics being what it is there is no chance worth mentioning that we will agree on "the right way" to do this problem. Apparently physicists will have stronger prior opinions about f_s than about μ_s. In the direction of further realism, the background and signal means may be written in terms of a product of a few quantities, or further developed into a sum of such terms. If this is done then there might be useful prior information about some of the terms in the products; at the same time one of the terms in the product or products leading to μ_s would be a quantity of great interest and physicist would not like any prior on this term to play any important role.

The Challenge

The idea is that someone will generate realistic data in the context of one of the low numbered Frameworks above, probably Framework 3, and we will make suggestions for methods for testing for the null hypothesis of no signal. Participants would likely be provided with information about λ_b, λ_s and the associated shapes and means in the form of training (Monte Carlo) samples. Notice, for instance, that when a specification such as in Frameworks 4 or 5 is parametric there would be a need for training samples from a variety of values of the parameter m.

Jargon

Here is my take on some jargon which will probably arise during the Workshop. We might want to add to this list from time to time; at the moment we certainly want to correct the definitions below:

High stats / low stats: describes large and small sample sizes (or perhaps large and small Fisher information).

Triggering: the hardware in the LHC discards nearly all events on the fly; retained events have, I suppose, triggered the acquisition and retention of data. The rate at which events are generated is so enormously large (on the order of 10^7 per second) that retaining all data is impossible.

Cuts: In the opening paragraph I described the process of reducing the number of events to a number like 10 to 1000. This process happens by setting down rules for discarding events; these are the cuts. They happen after the triggering process has already removed the vast majority of events. The cuts hopefully narrow the focus in the mark space to a region where the signal intensity might be distinguishable from the background intensity. Without the cuts the number of signal events is totally negligible relative to the number of background events.

Cross-section: when you shoot some particle at a target a bigger target is easier to hit. The cross-section of a target is essentially the probability of an event (usually an event of some specific interesting type) per unit area of target (and per incident particle?) The signal mean μ_s is proportional to the cross-section for the production of the particle of interest. This is the crucial term mentioned in Framework 5 above.