BC2a Document

M. T. Allen

This document is intended to serve as documentation of the entry into Banff Challenge 2a.

1 Outline of Method

We perform an extended, unbinned maximum-likelihood fit to estimate the probabality of a signal a given dataset, and the location of the signal should it exist. In the fit, we allow both the peak position and the number of signal events to float. We calculate the P-value for the null hypothesis by comparing a signal strength parameter to a large sample of generated datasets without signal events. We use the pyminuit (http://code.google.com/p/pyminuit/) implimentation of the Minuit (http://seal.web.cern.ch/seal/snapshot/work-packages/mathlibs/minuit/) minimization package to both minimize the negative log-likelihood and to determine the 68% CL.

2 Minimization proceedure

We perform an extended, unbinned maximum likelihood fit to each dataset. The likelihood function, $\mathscr{L}$, is defined as:

$$\mathscr{L} = \mathrm{e}^{-\left(n_{sig}+n_{bkg}\right)} \prod_{j=1}^{N} \left[n_{sig}\mathscr{P}_{sig}(x_j; E) + n_{bkg}\mathscr{P}_{bkg}(x_j)\right] \tag{1}$$

Were n_{sig} and n_{bkg} are the number of signal and background events, respectively, x_j are the input data, and $\mathscr{P}$ represents the probability distribution functions of the background and signal hypotheses. They are defined as:

$$\mathscr{P}_{sig}(x; E) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-E)^2}{2\sigma^2}}; \sigma = 0.03 \tag{2}$$

$$\mathscr{P}_{bkg}(x) = \frac{1}{\lambda} e^{-\lambda x}; \lambda = 10 \tag{3}$$

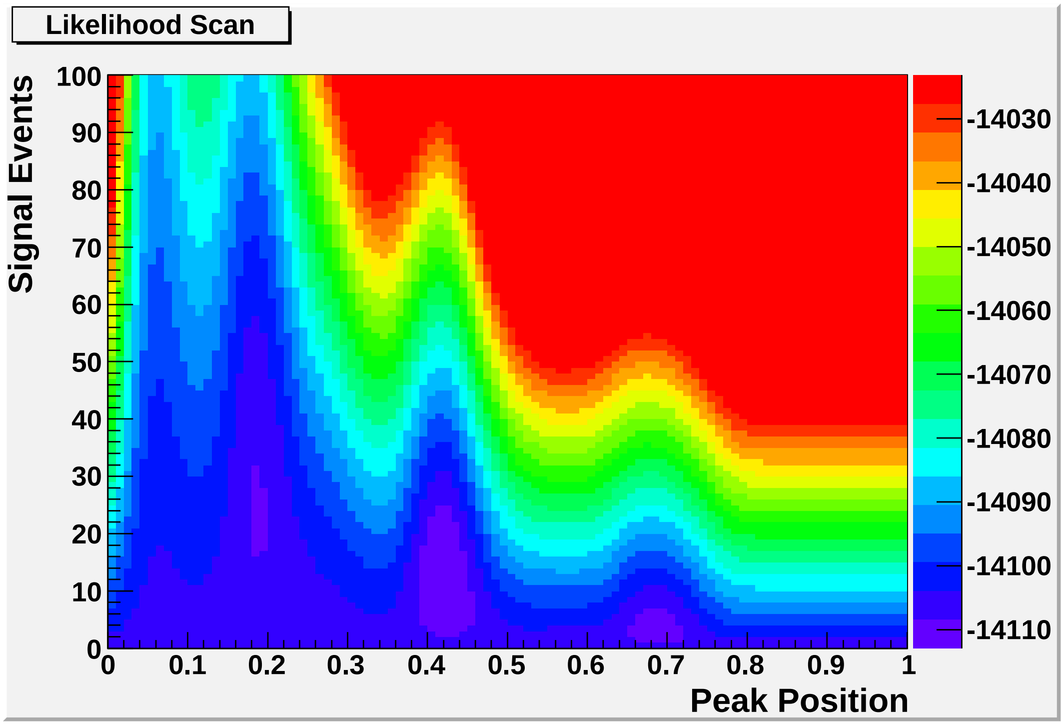

Figure 1: Likelihood scan of a typical dataset. The color scale represents $-\ln\mathscr{L}$. Several minima are visible.

The parameters left to float are n_{sig}, n_{bkg}, and E. In practice, $-\ln(\mathscr{L})$ often has several local minima. We choose as the result whichever minima is the least. Figure 1 shows a typical likelihood scan of the n_{sig}-E parameter space. Such scans are not particularly effective in finding the correct (deepest) minimum. It is quite common for local minima to be narrow, but computational concerns require a finite resolution. Instead, we repeat the fit with many sets initial parameters, and choose the set of parameters with the lowest minimum.

3 Signal and P-value Determination

In order to determine whether to claim a signal, we generate a series of datasets following the background distribution as deined in the BC2 problem description, and no signal events (hereafter called the "Null" set). For each dataset within the Null set we follow the minimization proceedure, and find the best fit to the data. We then calculate the value of the likelihood under the assumption of zero signal events. The difference between the value of the log likelihood with and without allowing for signal events represents the strength of the signal. We define a parameter δ as

$$\delta \equiv \ln \mathscr{L}(min) - \ln \mathscr{L}(n_{sig} = 0)). \quad (4)$$

The distribution of δ from the Null set is shown in Figure 2, and the normalized cumulative distribution, $CDF(\delta)$, is shown in Figure 3. The 99th percentile is $\delta_{99} = 11.292$, which gives

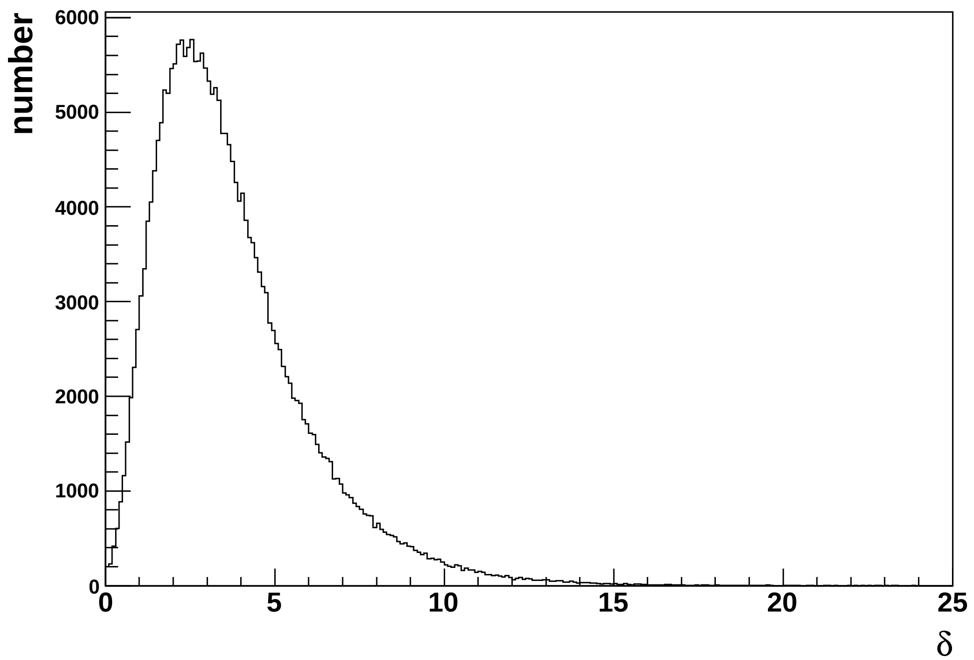

Figure 2: The δ distribution from fitting 250,000 mock datasets with no signal events.

us the criteria for claiming a signal. We estimate the error on δ_{99} using a bootstrap technique. Using a total of 250,000 toy datasets, and 1000 bootstrapped samples, the estimated error on δ_{99} is 0.042. We determine the P-value for a given dataset to be

$$P_i \equiv 1 - CDF(\delta_i), \tag{5}$$

where δ_i is the δ measured for a given dataset. In order to determine if δ_{99} is independent of peak position, we bin the Null set results into 20 E bins, and calculate the 50th, 75th, 90th, 95th, and 99th percentiles of the δ distributions. The results are in Figure 5. The value of the 99th percentile is mostly independent of peak position.

4 68% CL determination

Should a dataset pass the δ requiremnent above, we use the best fit values for the peak position, E, and the number of signal events, n_{sig} as the central values. We use the MINOS asymetric errors from the Minuit fit package to determine one σ errors. A small multiplicative adjustment is applied to turn it into the 68% CL. Our method produces an estimated number of signal events, however the problem description defines the signal as D times an un normalized Gaussian. To turn the number of signal events into D, we simply divide by the Gaussian integral.

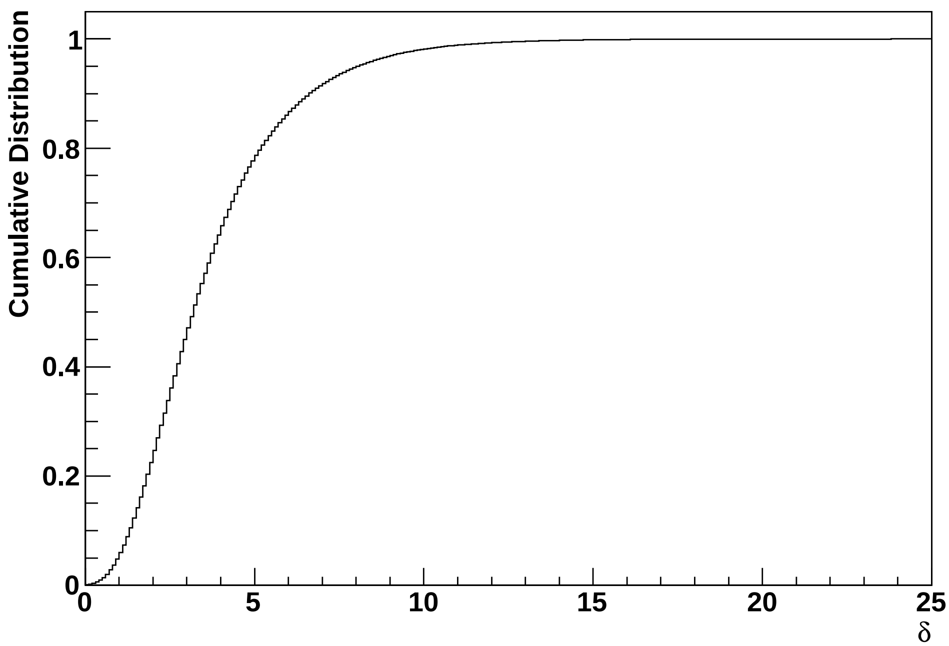

Figure 3: The normalized cumulative distribution of δ from the mock dataset without signal events. The curve is 0.99 at $\delta = 11.292$.

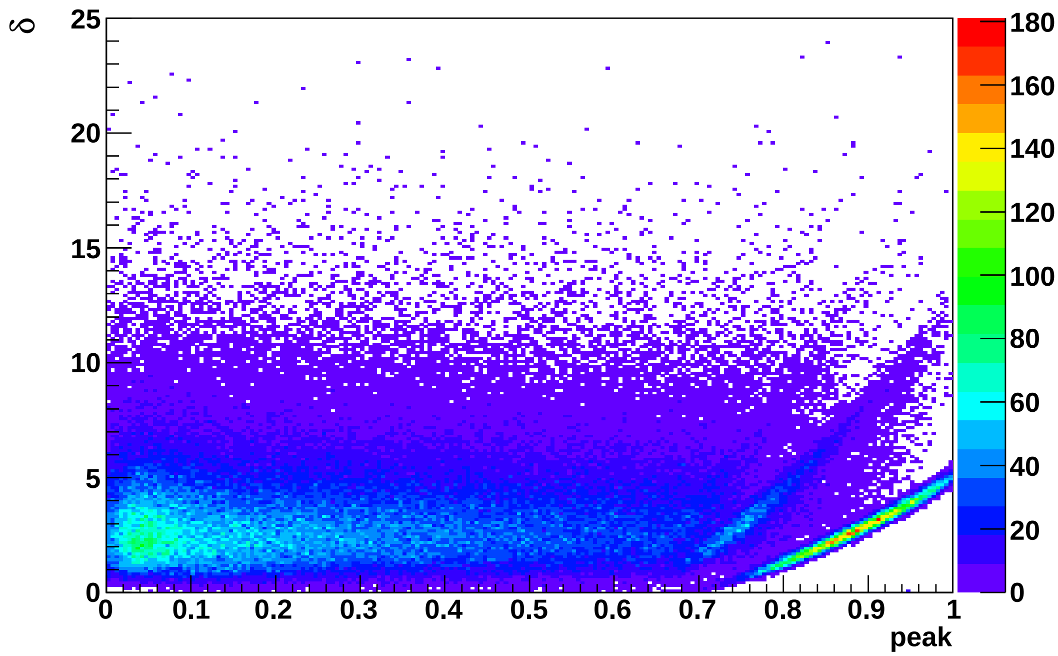

Figure 4: Distribution of δ as a function of the peak position. The structure at a peak value of 0.9 represents individual events on top of very low expected backgrounds.

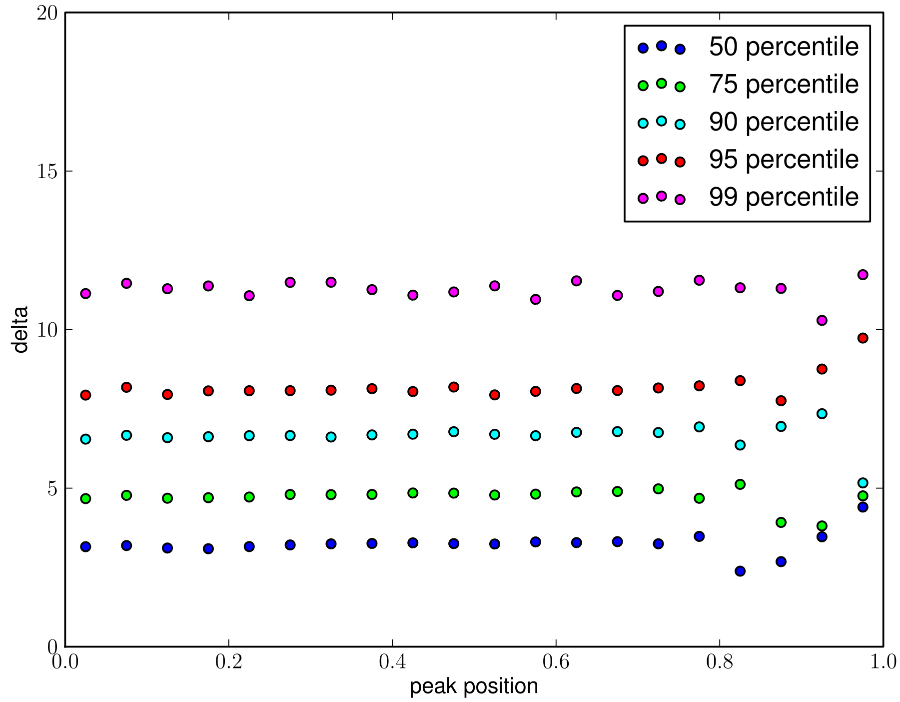

Figure 5: The 50, 75, 90, 95, and 99th percentile δ values, as a function of peak position.

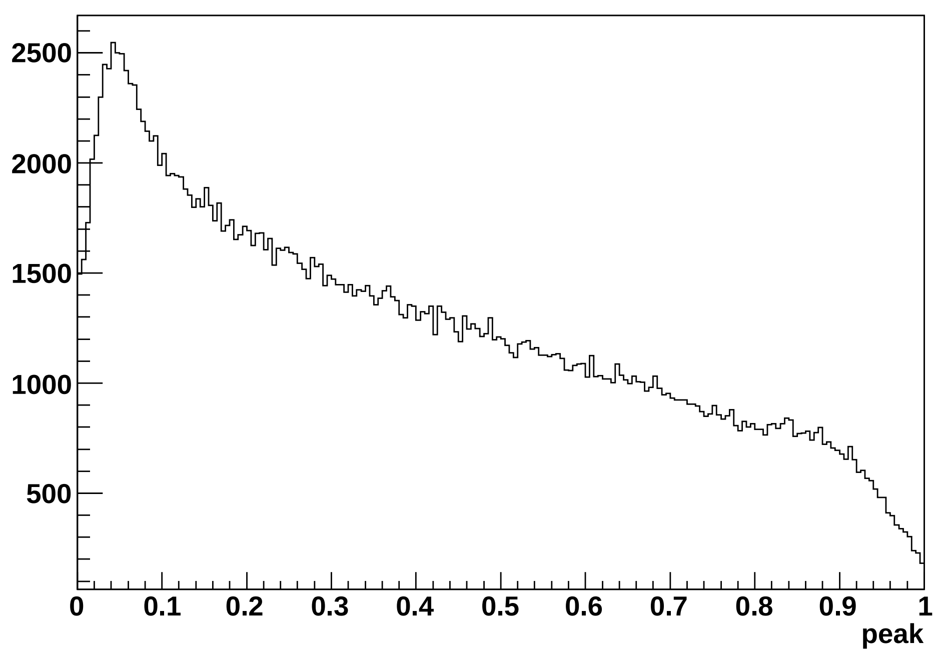

Figure 6: Distribution of best fit peak values for the no signal datasets.

D	E	Rate
1010.0	0.1	0.385
137.0	0.5	0.486
18.0	0.9	0.187

Table 1: Results of the power/sensitivity test.

5 Predicted Results

We have generated a series of mock datasets that are meant to simulate the datasets given in the problem. For each dataset we blindly apply our proceedure and compare our results to the known input values. The datasets are generated by randomly generating 1000 ± 100 background events. Half of the datasets get no signal events. For the other half we uniformly generate a peak position between 0 and 1. The number of signal events is generated from a Poisson distribution with a mean of $N_s = 1000 * r * \left(1 + \int_{E-0.03}^{E+0.03} dx\, e^{-10x}\right)$, where r is uniformly distributed from 1 to 5.

We report the results below:

Fraction of datasets with no signal and signal claimed	0.0099	(1183/119069)
Fraction of datasets with signal and signal claimed	0.7437	(88594/119126)
datasets with signal and signal claimed:		
Fraction of sets where E is within the 68% CL	0.674	(54419/80791)
Fraction of sets where n_{sig} is within the 68% CL	0.686	(55453/80791)

6 Power/Sensitivity

We test the power/sensitivity of this technique by generating 40,000 toy samples for each of the three cases given, and calculating the fraction of toy sets that we would claim a signal. The results are in Table 6:

7 Appendix

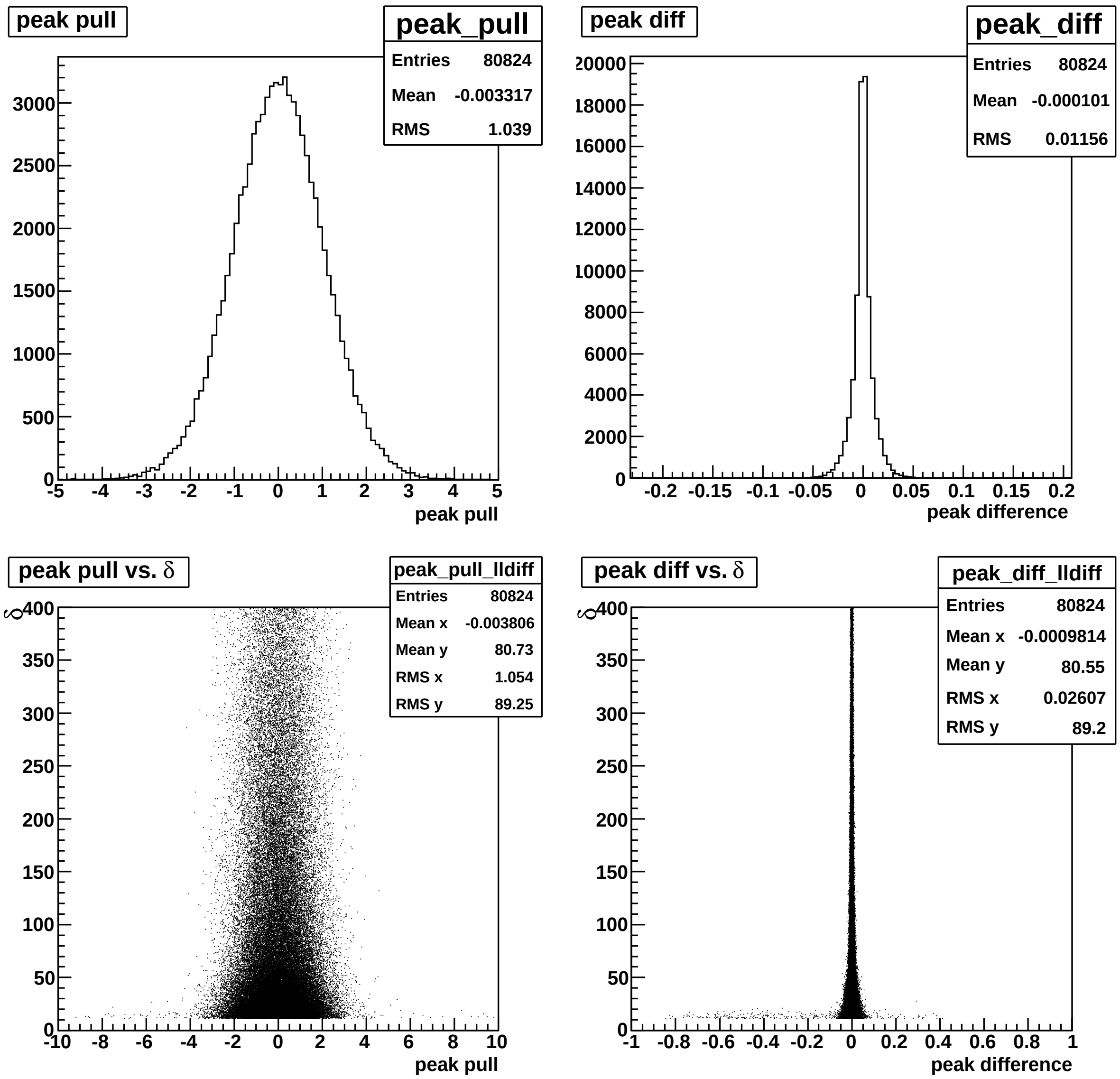

Figure 7: Diagnostic plots for the measurement of the peak position. The upper left hand plot is the peak pull, using the asymetric errors when possible, while the bottom left is the peak pull distribution plotted against δ. The upper right hand plot is the difference between the peak position and measured peak position, and the bottom right shows the difference between the peak position and measured peak position plotted against δ. The large scatter at low δ in the lower right hand plot is a result of the existence of a signal, but a larger background fluctuation, and thus the peak is poorly estimated.

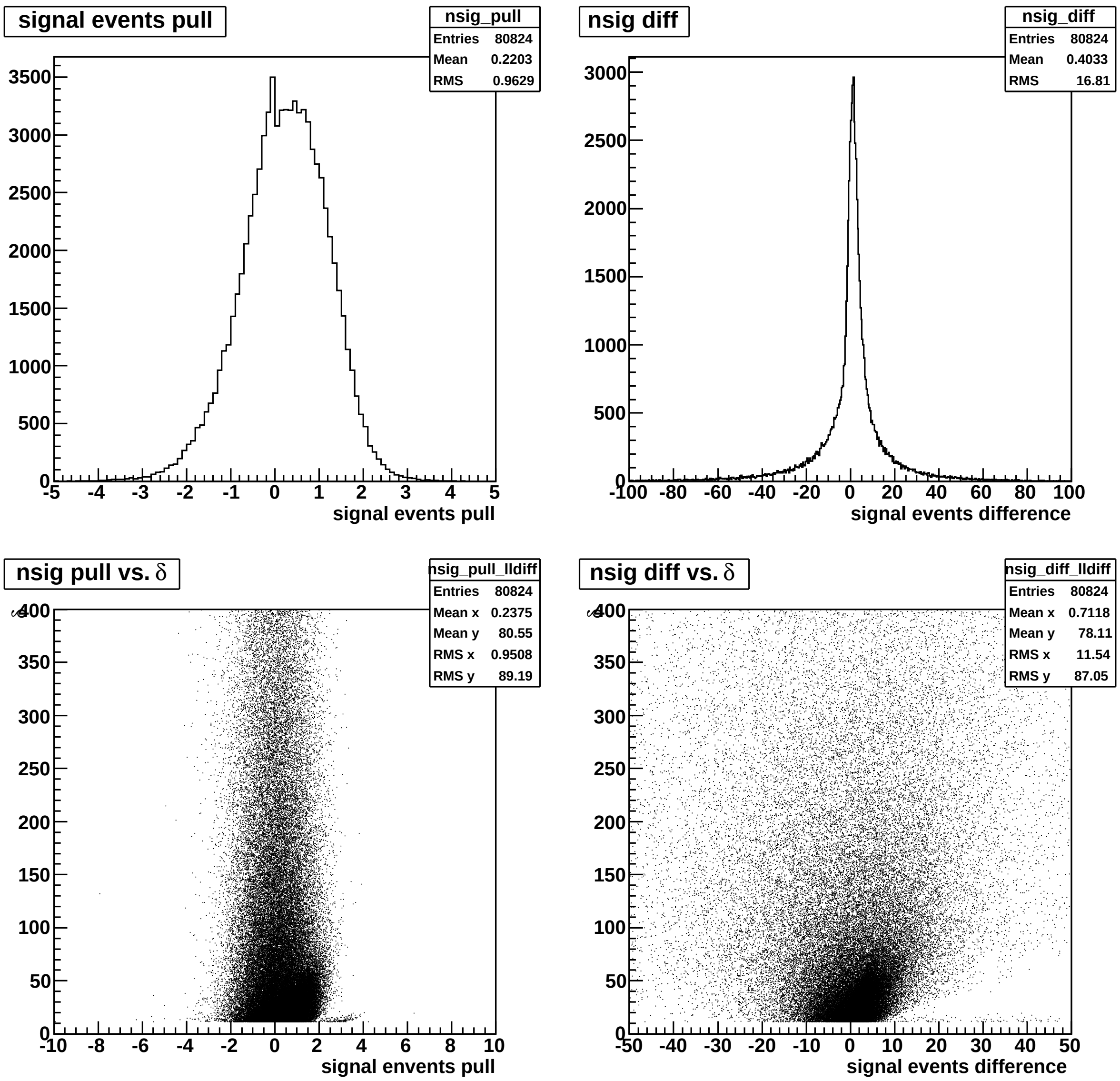

Figure 8: The upper left hand plot is the signal events pull, using the asymetric errors when possible, while the bottom left is the signal events pull distribution plotted against δ. The upper right hand plot is the difference between the signal events and measured signal events, and the bottom right shows the difference between the peak position and measured signal events plotted against δ. The bias in the signal events pull distribution represents dendency to pick out as the appropriate minimum an upward fluctuation in the number of events. In other words, for a given trial, should the number of signal events fluctuate down, we don't see the signal, and that trial doesn't enter the plot, while an upward fluctuation is much more likely to be identified and so we get a bias.