Banff Challenge 2: Posing the Problem

Wade Fisher

Michigan State University

Tom Junk (filling in for Wade)

Fermi National Accelerator Laboratory

BIRS Workshop:

Statistical issues relevant to significance of discovery claims

BIRS

Outline

- × Some background on the challenge problem

A bit of particle physics: theory & experiment

Accessing parameters of interest

Classification schemes

Systematic uncertainties

Statistical tests

- × Banff Challenge 2

The problems

Wade's results

Resources

✖ There have been many excellent talks on the use of statistical methods in high energy physics (particle physics)

✖ PHYSTAT:

http://www-conf.slac.stanford.edu/phystat2003/

http://www.physics.ox.ac.uk/phystat05/

http://phystat-lhc.web.cern.ch/phystat-lhc/

✖ Previous BIRS conferences

✖ In particular, Tom Junk and Kyle Cranmer have given several excellent overview talks from which I drew some inspiration.

  Cranmer:http://phystat-lhc.web.cern.ch/phystat-lhc/program.html

  Junk: http://temple.birs.ca/~06w5054/tom-junk.pdf

Banff Challenge 2

- ✖ We are here discussing "Statistical issues related to significance of discovery claims"

- ✖ I am a particle physicist, so my mind is naturally attracted to the discovery potential of the Large Hadron Collider (LHC)

- ✖ Thus, the 2nd Banff Challenge is a mock-up of problems we expect to encounter at the LHC.

  - ✖ We will benefit from novel interpretations of the problem.
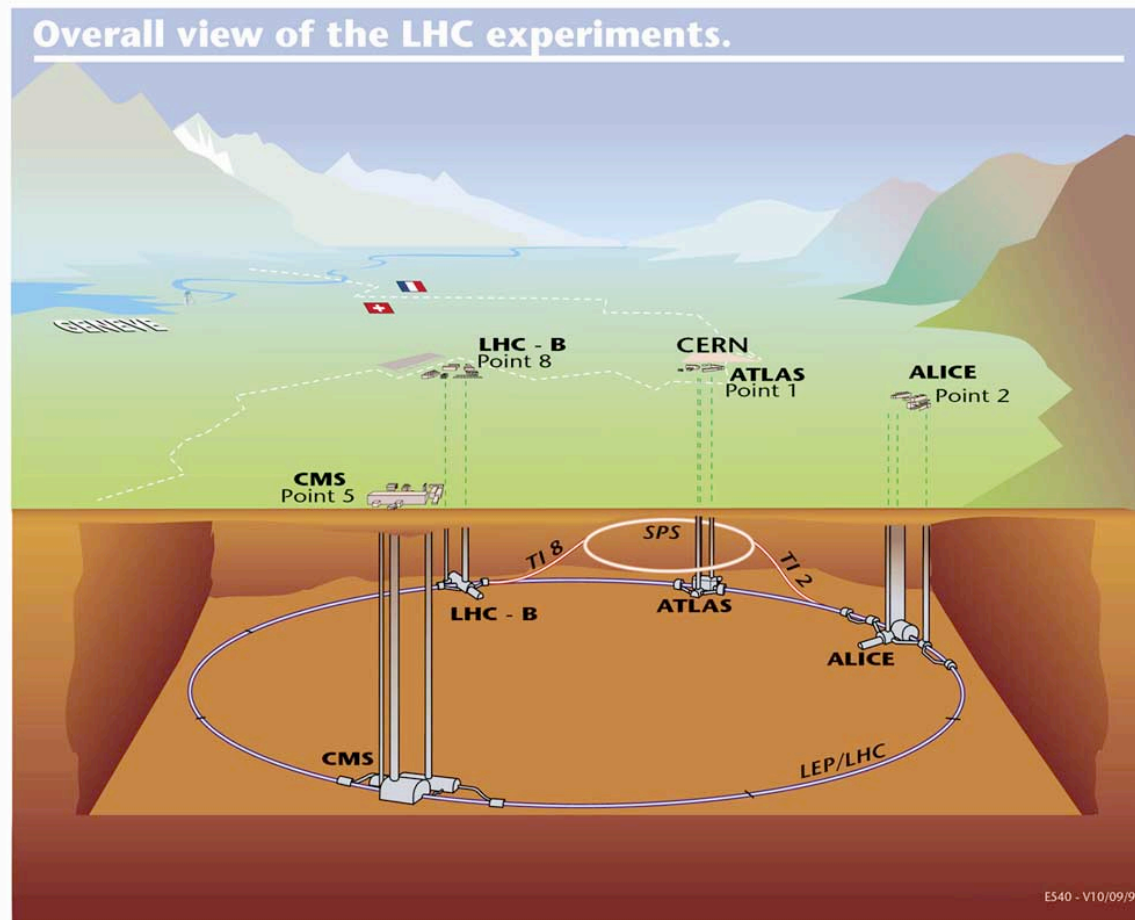
The Large Hadron Collider

Located at the European Center for Nuclear Research (CERN) outside Geneva

26 km in circumference

Collides bunches of protons at a rate of 40 MHz

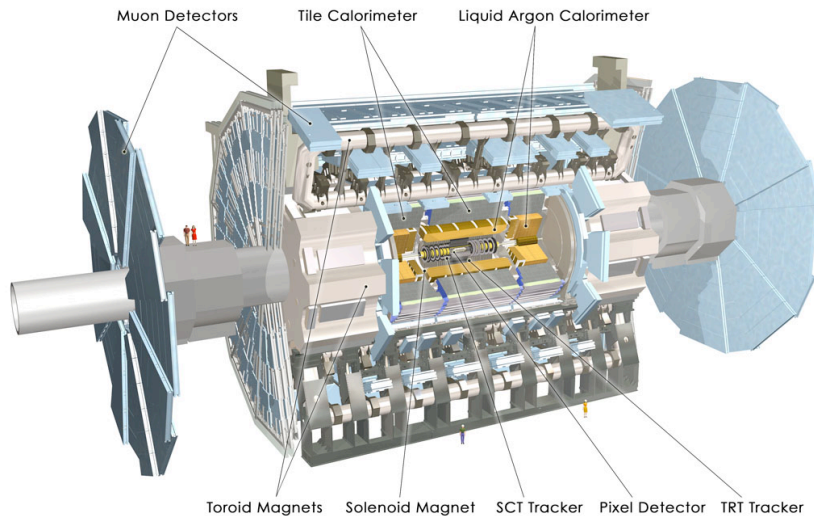Particles are accelerated to an energy of 7 TeV (proton rest mass is ~1 GeV)
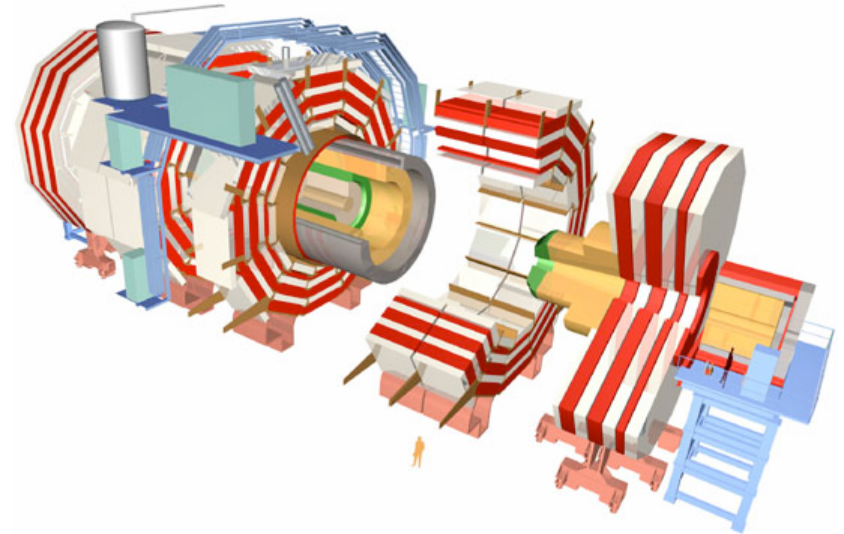
Center-of-mass energy 14 TeV

Overall view of the LHC experiments.

Experimental Appartatus

ATLAS
× Lenth   = ~40m
× Radius = ~10m
× Weight = ~7000 tons

CMS
× Lenth   = ~21.5 m
× Radius = ~7.5m
× Weight = ~12,500 tons

ATLAS and CMS are multi-purpose particle detectors

Designed to be able to address a wide range of potential physics signals

Collides bunches of protons at a rate of 40 MHz

Broad discovery potential will need to be addressed with well-designed statistical methods.
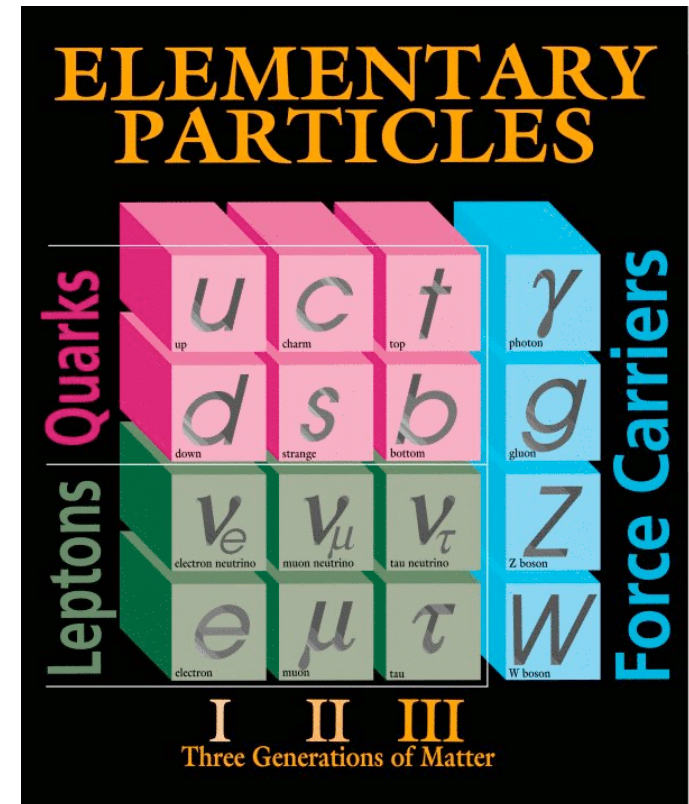
Particle Physics

The 'Standard Model' (SM) is the foundation of modern particle physics

Describes the fundamental building blocks of matter and their interactions

Six quarks and **six leptons** arranged in three generations of matter

Electroweak forces mediated by four electroweak bosons ($\boxed{\gamma}$, $W^{\pm}$, Z^{0})

Strong forces amongst quarks and nucleons mediated by gluons (**g**)

ELEMENTARY PARTICLES

I II III
Three Generations of Matter

Fermilab 95-759

Particle Physics

The 'Standard Model' (SM) is the foundation of modern particle physics

The SM is based on the theoretical formalism of Quantum Field Theory & allows us to predict:

The probability for an interaction to occur

✘   Called a "cross section"

Kinematical distributions for observables

✘   Particle directions, relative angles, masses, etc.

✘   Typicall denoted *f (x)*

We arrange opportunities for interesting particle interactions using particle accelerators (man made or cosmological origins)

Using our experimental apparatus, we collect data by sampling these interactions stochastically.

The number of observed events *N* is Poisson distributed.

N is the sum of the possible contributing processes.
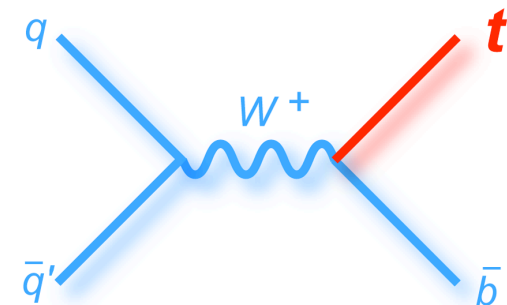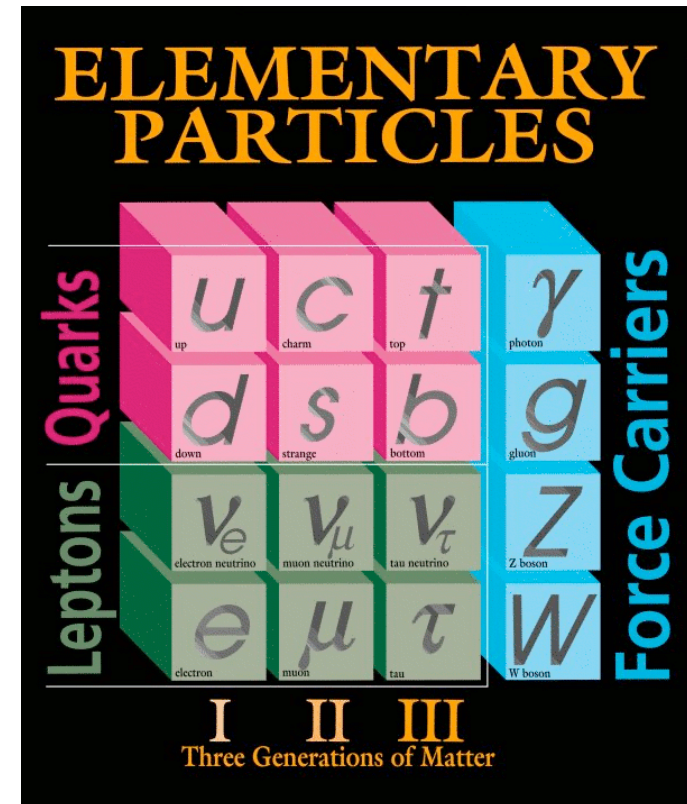
Testing Hypotheses

We have been testing the SM for many years

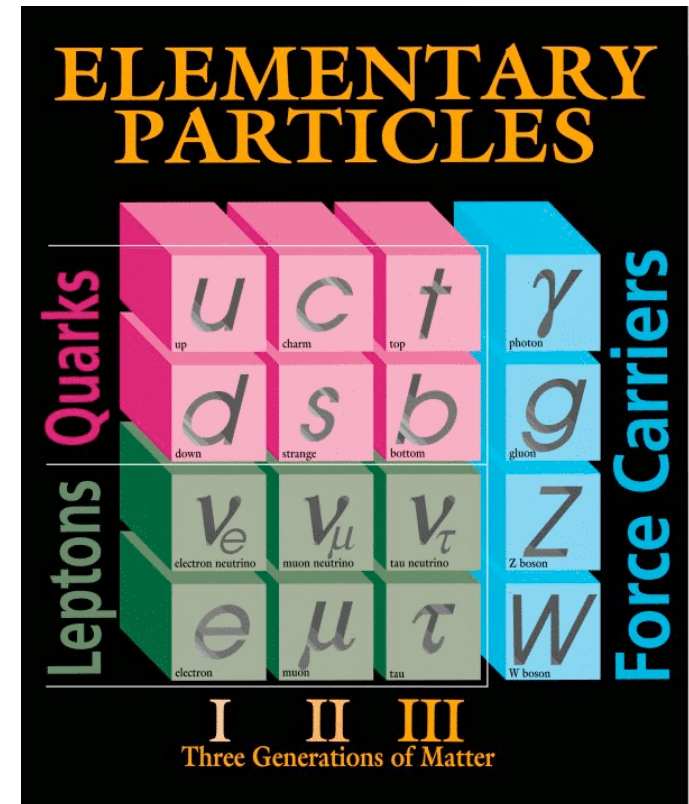A wide range of physics is known and can be predicted with great precision.

A major goal is to uncover signs of new physics

When we search for signs of new physics, the well-tested processes are our NULL hypothesis (H0)

Often referred to as the background-only hypothesis.

The expected rate for background denoted as **b**

$$L(x \mid H0) = Poiss(n \mid b) \prod_{j}^{N} f_{b}(x_{j})$$

Testing Hypotheses

For many reasons, we believe the SM is incomplete and expect some indication of new physics as we access higher energies (masses)

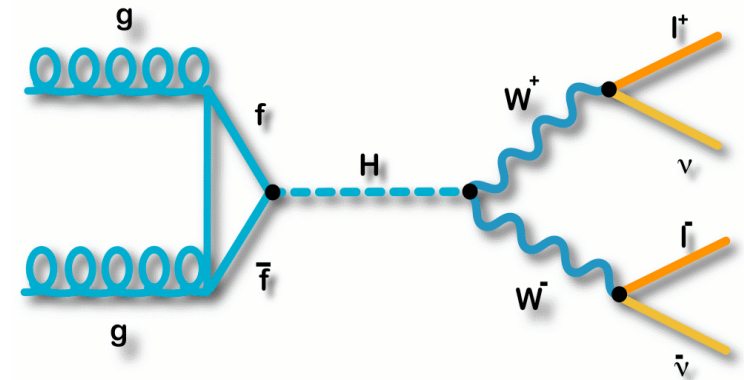New particles or interactions commonly referred to as the "signal"

Can generally be simulated as well as the "background" processes

Generally assume that signal will add to background, but this is not always the case!

Simulations/calculations predict the rate and probability densities

Denote signal rate as **s**

Probability density **f$_S$(x)**

$$L(x \mid H1) = Poiss(n \mid s+b) \prod_j^N \frac{sf_s(x_j) + bf_b(x_j)}{s+b}$$

Parameter of Interest

The parameter of interest can vary

Commonly the cross section (rate) for the new physics signal

Could be a particle mass (or both!)

Can be something more nuanced, like the size of an interference effect.

We generally try to parametrize our model parameters such that we can identify the parameter of interest as the signal cross section and potentially test many possibilities.

The parameter of interest is generally not accessible via all observables

What we observe in the detector must be reconstructed in order to access a fundamental property of the particle production process.

We are limited to relatively simple methods to extract information: Momentum/Energy/Charge conservation.

From this and our experimental measurements, we infer properties on an event-by-event basis.
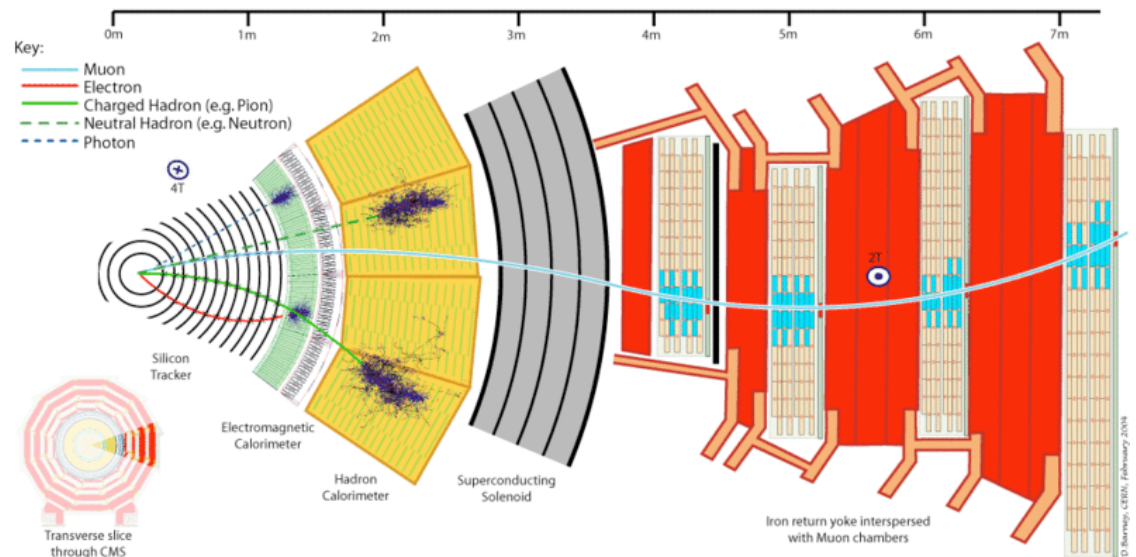
Systematic Uncertainties

In HEP jargon, uncertainties on nuisance parameters are called 'Systematic Uncertainties'

Associated with every aspect of a measurement:

Energy & momentum measurements

Integrated luminosity measurements

Theoretical predictions for cross sections (event rates)

Theoretical predictions for probability densities

Our observables are generally a complicated function of nuisance parameters

We rarely know the true functional form

We thus must describe our nuisance parameter model as the empirical change in observable found by varying the nuisance parameter.

$$O_{nominal} = \alpha \times F(\theta_{nom})$$

$$O_{\pm 1\sigma} = \alpha \times F(\theta_{nom} \pm \delta\theta)$$

The 2nd Banff Challenge

BIRS

As noted, the 2nd Banff Challenge is modeled after what one may expect from LHC search results

An experiment collects data for some period of time, corresponding in some integrated luminosity value.   I = instantaneous luminosity

$$\mathcal{L} = \int_{0}^{t_{final}} I dt$$

The number of observed data events N depends on what processes can be generated, how events are selected, etc. We can simulate this using Monte Carlo simulations. For this we need a normalization and a probability distribution.

Normalization: To predict the rate of a specific process we need to know

   1) The probability for the process to occur (the 'cross section'): β

   2) The number of opportunities the process had to occur (luminosity): $\mathcal{L}$

   3) The fraction of the total possible occurrences our experimental apparatus and selection will select (the efficiency): ε

$$N_{events} = \mathcal{L} \times \beta \times \varepsilon$$

As noted, the 2nd Banff Challenge is modeled after what one may expect from LHC search results

An experiment collects data for some period of time, corresponding in some integrated luminosity value.   I = instantaneous luminosity

$$\mathcal{L} = \int_{0}^{t_{final}} I dt$$

The number of observed data events N depends on what processes can be generated, how events are selected, etc. We can simulate this using Monte Carlo simulations. For this we need a normalization and a probability density.

Probability density: The simulation tells us about the kinematics of each event we generate.

The accumulation of such events describes the intensity λ(**x**) and, thus, the probability density **f(x)**

$$f(x) = \frac{\lambda(x)}{\int \lambda(x) dx}$$

The 2nd Banff Challenge

BIRS

There are three challenge problems. Each problem consists of:

1) A distribution of observed data events, each characterized by some observable.

2) 500k generated signal and background events, used to define $\lambda^S(x)$ and $\lambda^B(x)$

3) Instructions for normalization: $\beta \mathscr{L} \varepsilon$, provided for background and signal

4) Definitions of nuisance parameters and their uncertainties.

   A) Uncertainty on the true integrated luminosity (±10% -- correlated for signal & bkgd)
   B) Uncertainty on the signal and background efficiencies (±10% each, uncorrelated)
   C) Uncertainty on the background cross section (±10%)
   D) Uncertainty on the signal and background probability densities (shapes, see later)

   For simplicity, we assign a Gaussian prior for each nuisance parameter.
   The mean value (μ) is the nominally specified value for each parameter.
   The width (σ) is provided for each parameter separately.

Challenge Problem 1

Challenge 1: Simulated "Multivariate Classifier"

Exponentially falling background prediction.

Exponentially rising signal prediction.

$\mathscr{L}$ = **100 inverse femtobarns**

$\beta_{background}$ = **2000 femtobarns**

$\varepsilon_{background}$ = **0.05**

β_{signal} = **4.2 femtobarns**

ε_{signal} = **0.50**

Predicted Yields:

$n_{background}$=10000

n_{signal}=210

n_{data}=9815

Data + Normalized Predictions
Histogrammed with 50 bins
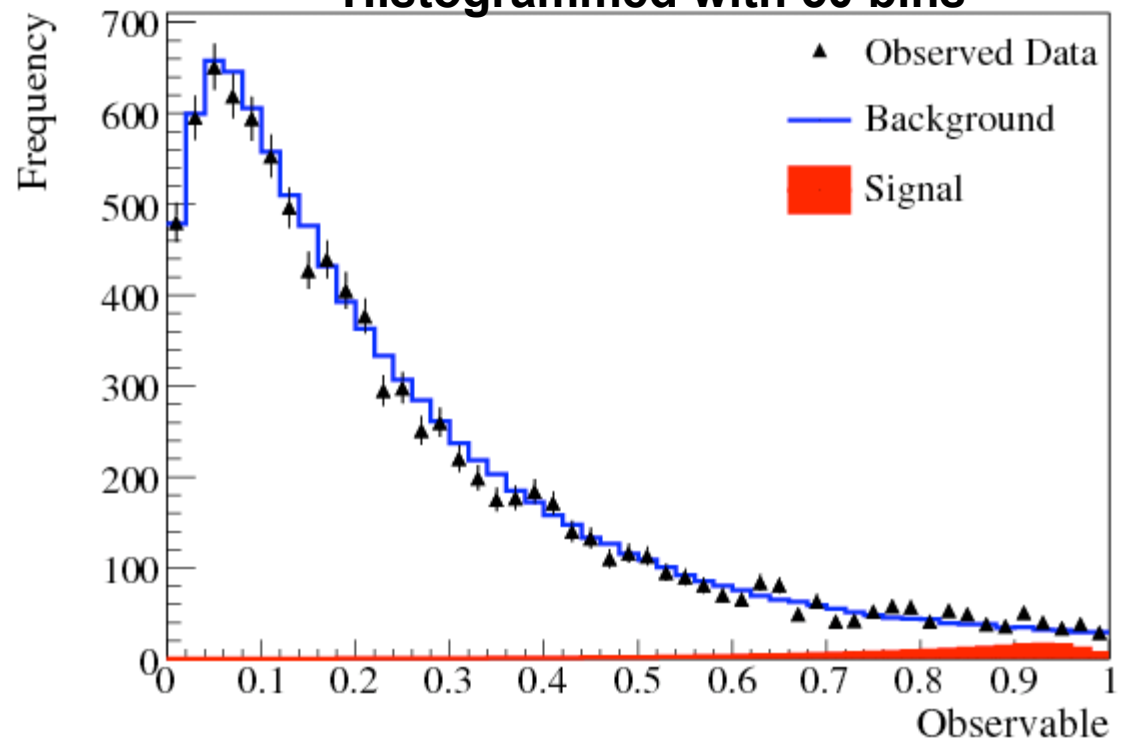
▲ Observed Data
— Background
■ Signal

Challenge Problem 1

Challenge 1: Simulated "Multivariate Classifier"

Exponentially falling background prediction.

Exponentially rising signal prediction.

$\mathscr{L}$ **= 100 inverse femtobarns**

$\beta_{background}$ **= 2000 femtobarns**

$\varepsilon_{background}$ **= 0.05**

β_{signal} **= 4.2 femtobarns**

ε_{signal} **= 0.50**

predicted yields:

$n_{background}$=10000
n_{signal}=210
n_{data}=9815

Probability Densities

Background-subtracted Data
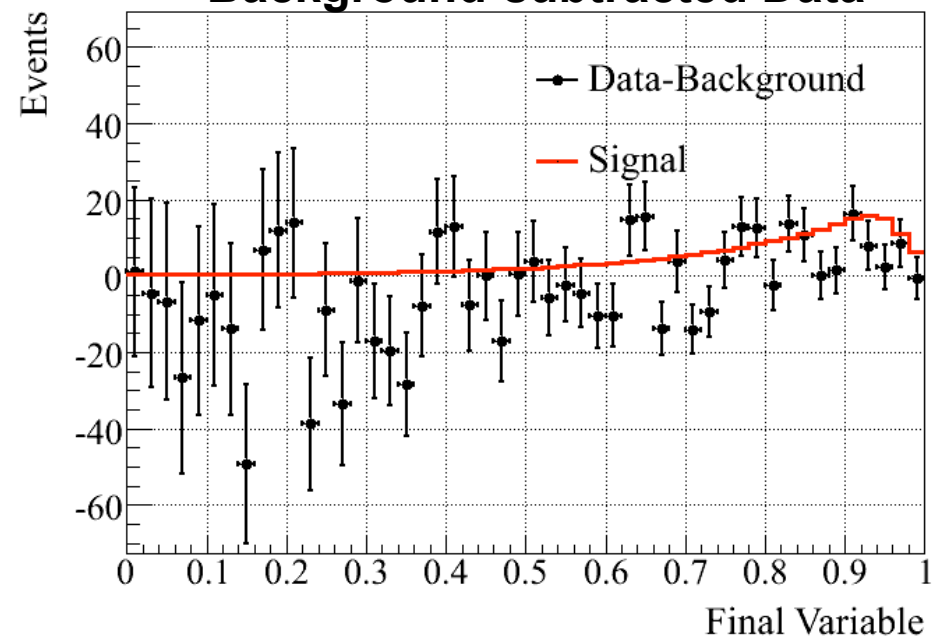
Challenge 1: Uncertainties on probability distributions
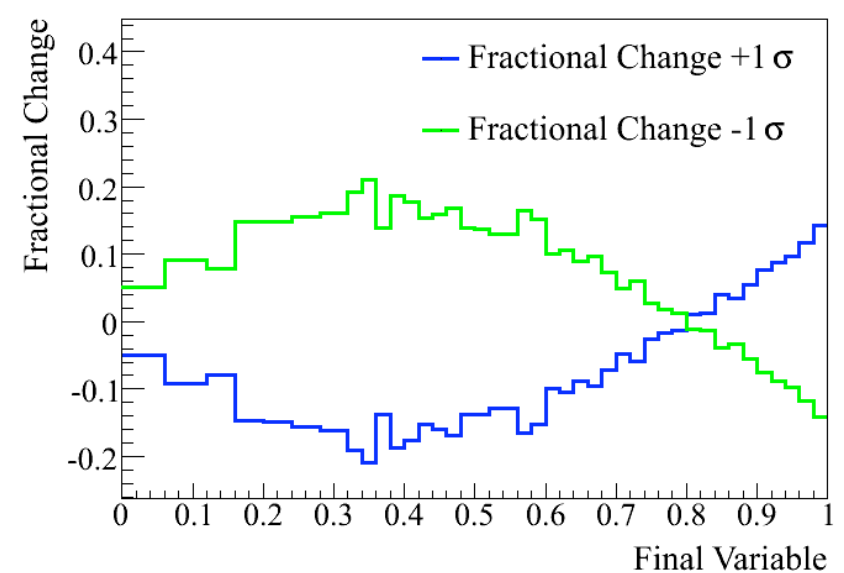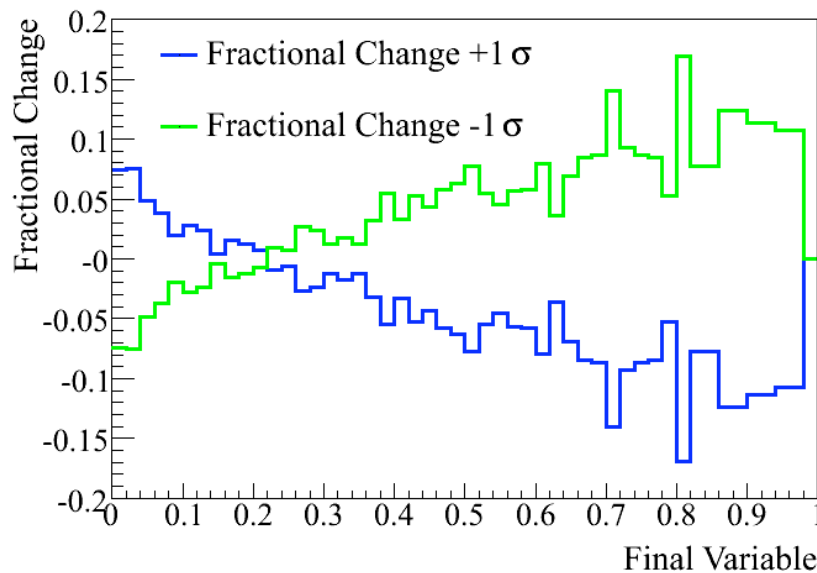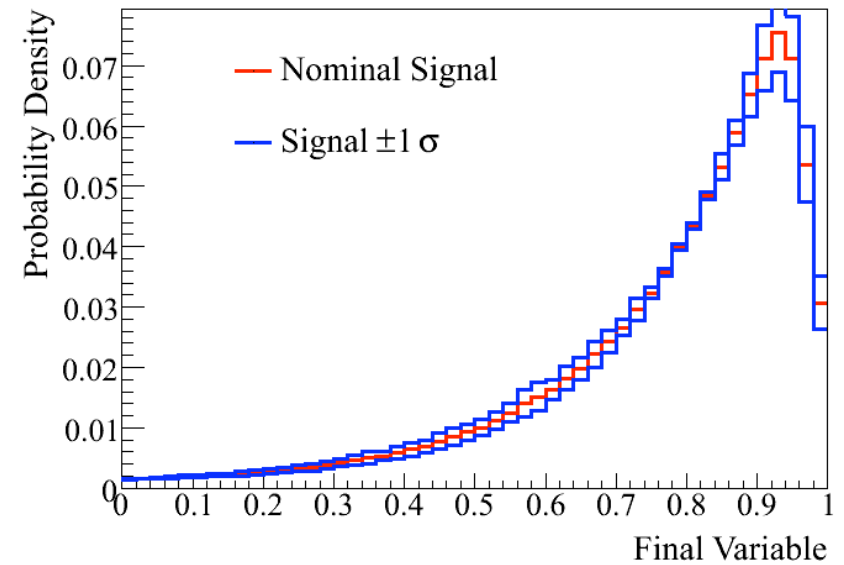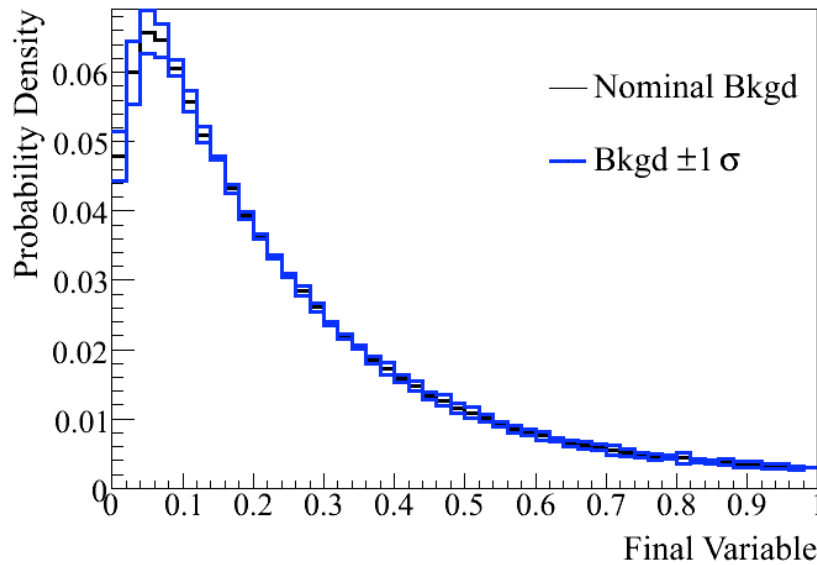
Challenge Problem 2

Challenge 2: Simulated Invariant Mass "Bump Hunt"

Exponentially falling background prediction.

Gaussian signal peak

$\mathscr{L}$ = 100 inverse femtobarns

$\beta_{background}$ = 2000 femtobarns

$\varepsilon_{background}$ = 0.05

β_{signal} = 4.0 femtobarns

ε_{signal} = 0.50

Predicted Yields:

$n_{background}$=10000
n_{signal}=200

n_{data}=9843

**Data + Normalized Predictions
Histogrammed with 50 bins
(well, 51 bins with one not shown..)**
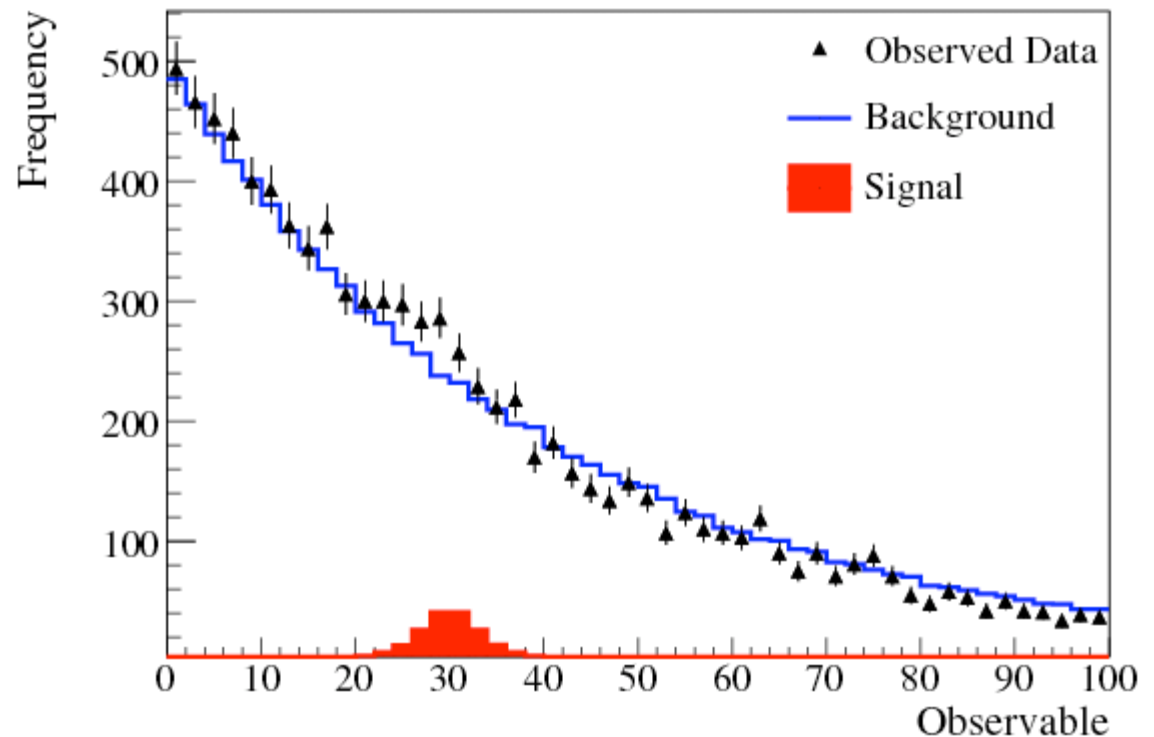
Challenge Problem 2

Challenge 2: Simulated Invariant Mass "Bump Hunt"

Exponentially falling background prediction.

Gaussian signal peak

$\mathscr{L}$ = 100 inverse femtobarns

$\beta_{background}$ = 2000 femtobarns

$\varepsilon_{background}$ = 0.05

β_{signal} = 4.0 femtobarns

ε_{signal} = 0.50

Predicted Yields:

$n_{background}$=10000
n_{signal}=200
n_{data}=9843
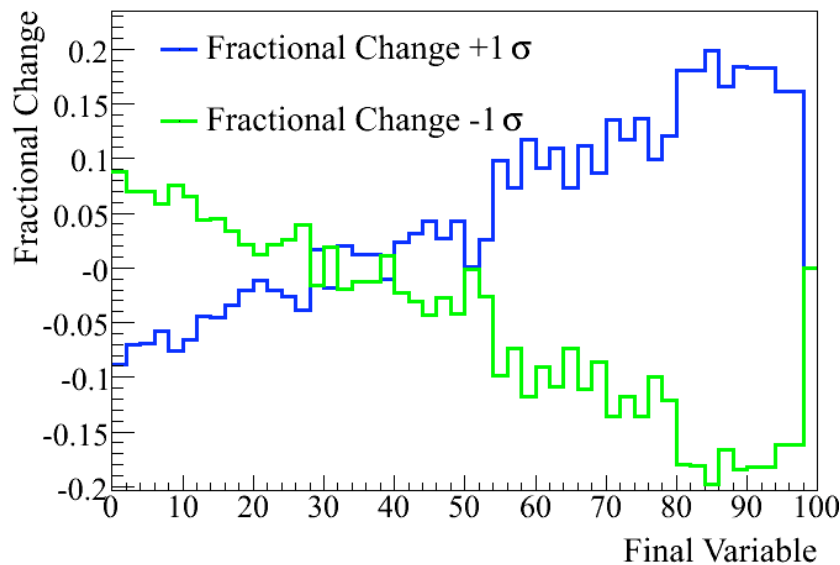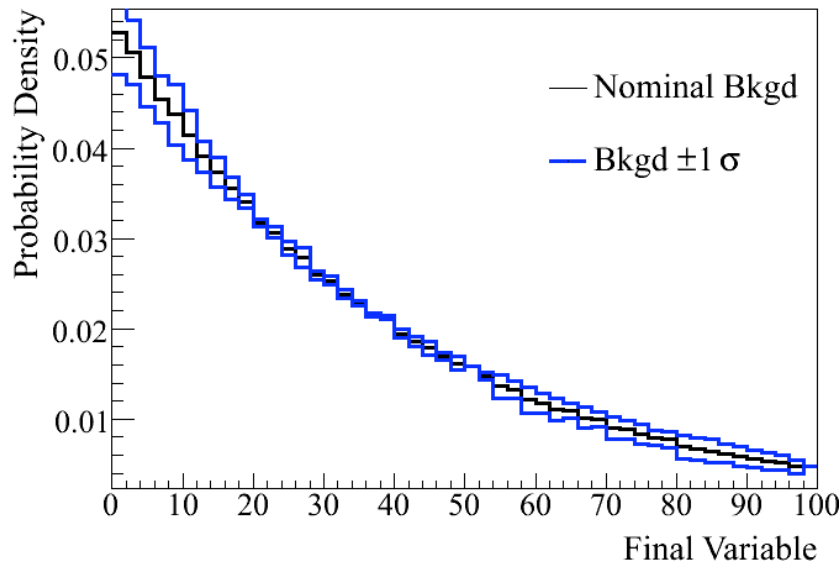

Probability Densities
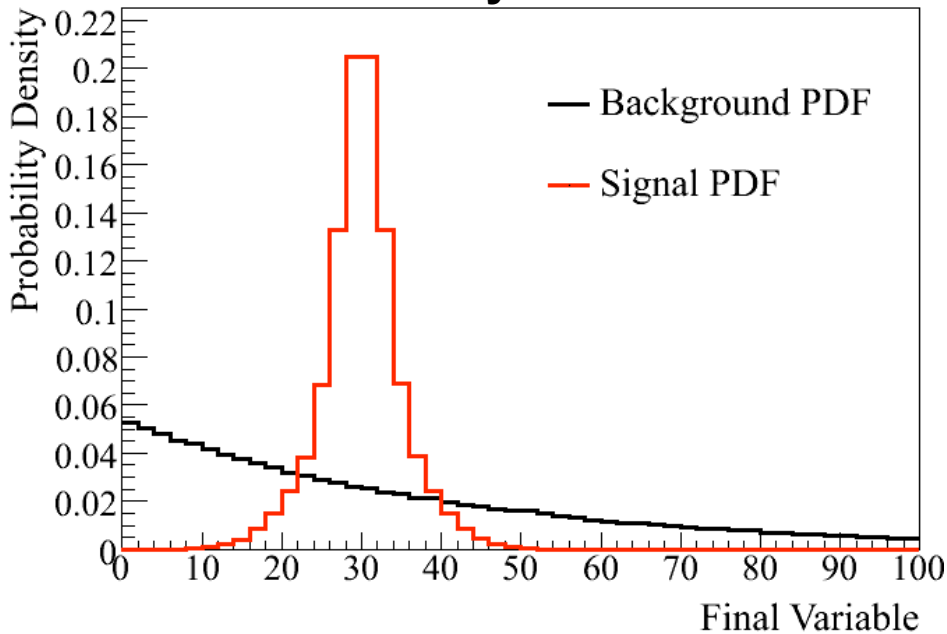
Background-subtracted Data
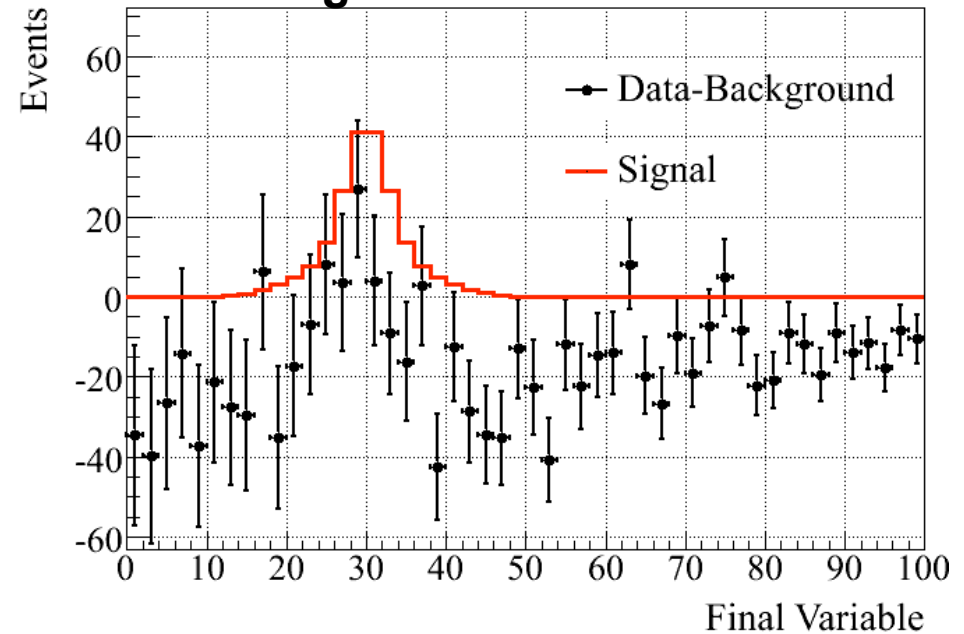
Challenge Problem 2

Challenge 2: Uncertainties on probability distributions
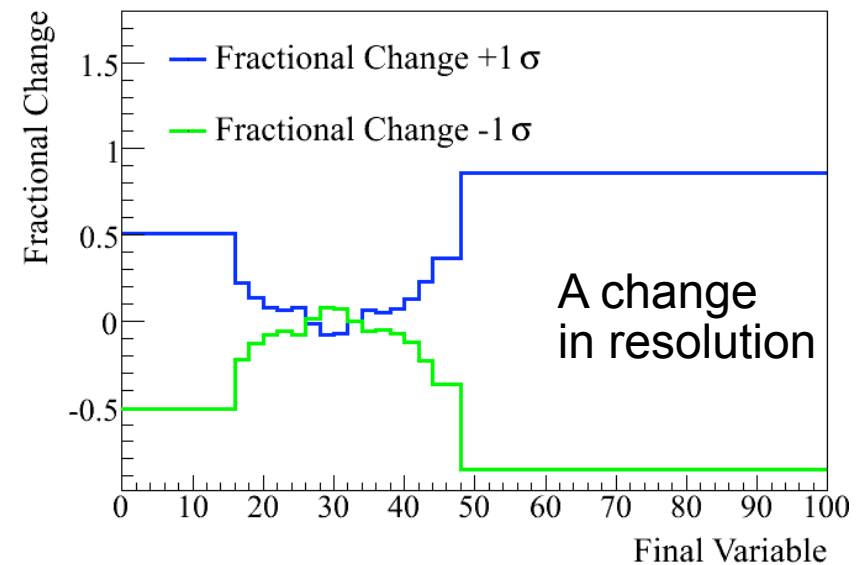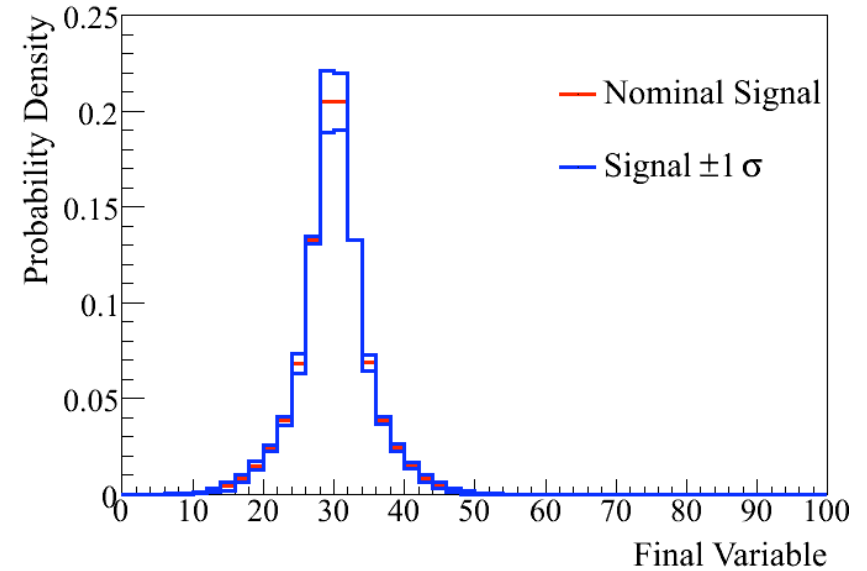
A change
in resolution

Challenge Problem 3

Challenge 3: Low-statistics, irreducible background

Background and signal exhibit nearly identical probability densities

Lower statistics search

$\mathscr{L}$ = **10 inverse femtobarns**

$\beta_{background}$ = **800 femtobarns**

$\varepsilon_{background}$ = **0.01**

β_{signal} = **18.0 femtobarns**

ε_{signal} = **0.40**

$n_{background}$=80
n_{signal}=72
n_{data}= 134

**Data + Normalized Predictions
Histogrammed with 50 bins**
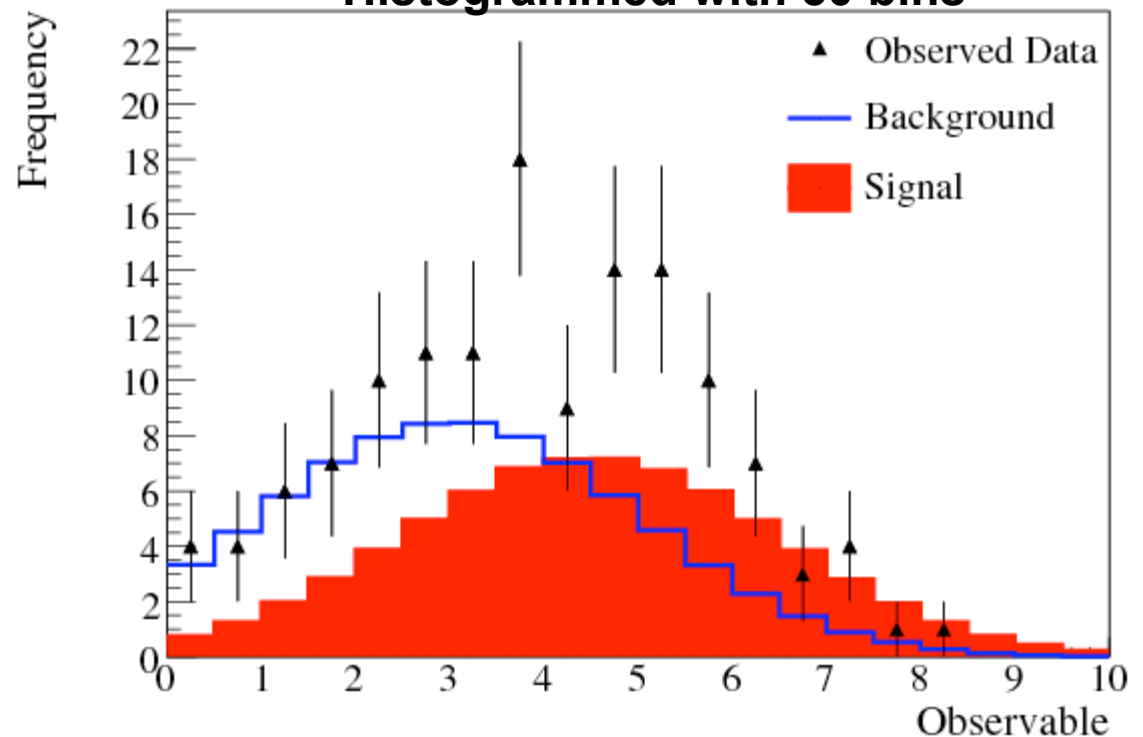
Challenge Problem 3

Challenge 3: Low-statistics, irreducible background

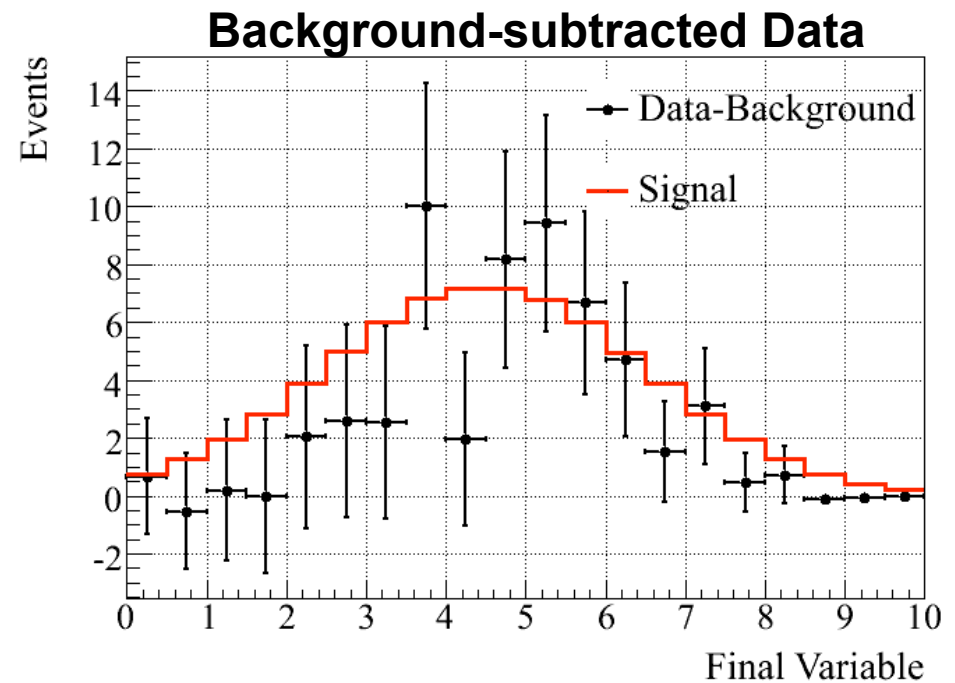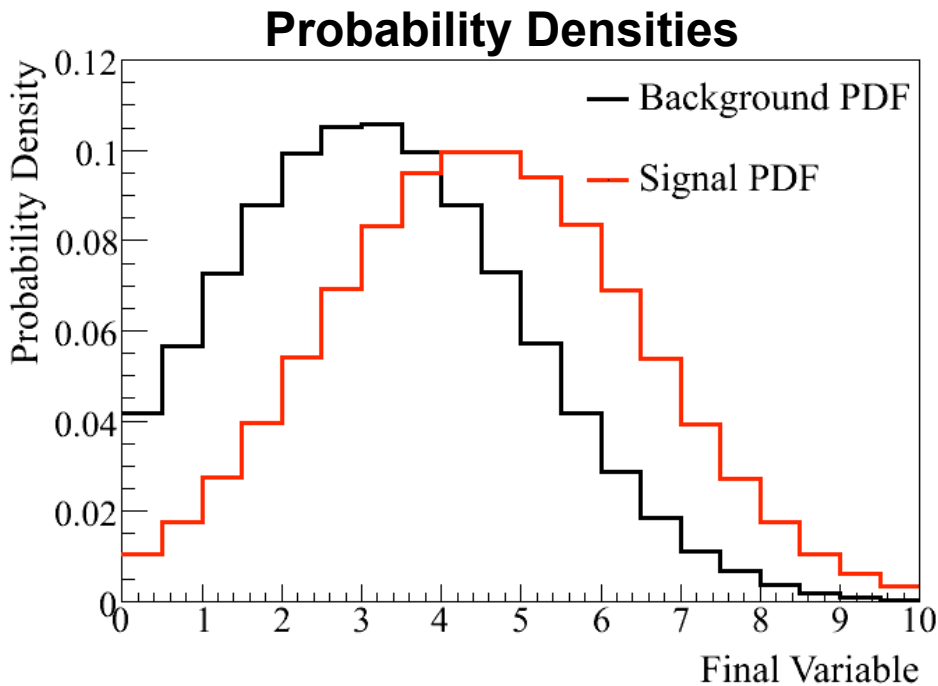Background and signal exhibit nearly identical probability densities

Lower statistics search

$\mathcal{L}$ = **10 inverse femtobarns**

$\beta_{background}$ = **800 femtobarns**

$\epsilon_{background}$ = **0.01**

β_{signal} = **18.0 femtobarns**

ϵ_{signal} = **0.40**

Probability Densities

Background-subtracted Data

Challenge 3: Uncertainties on probability distributions
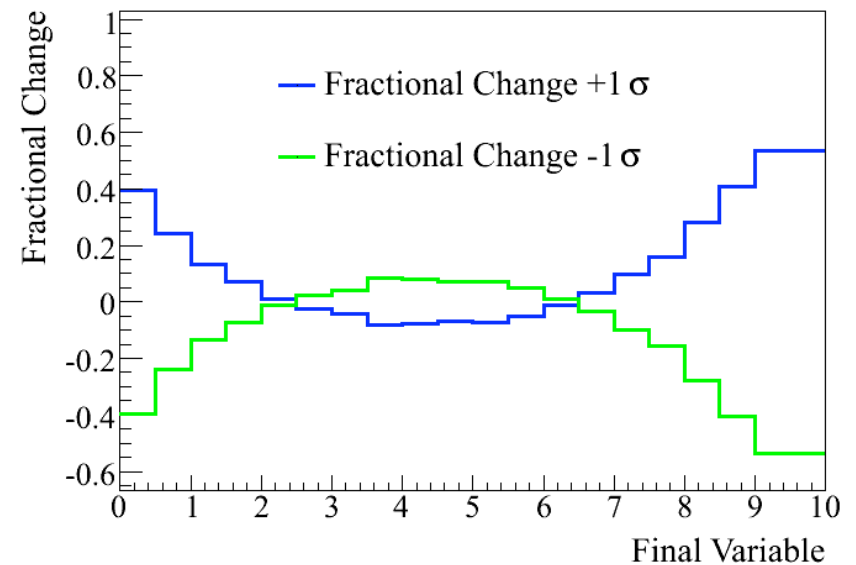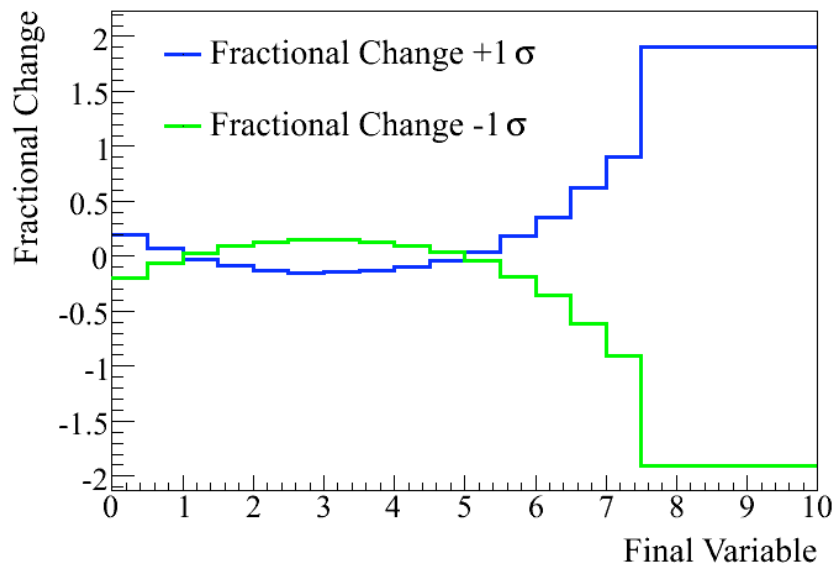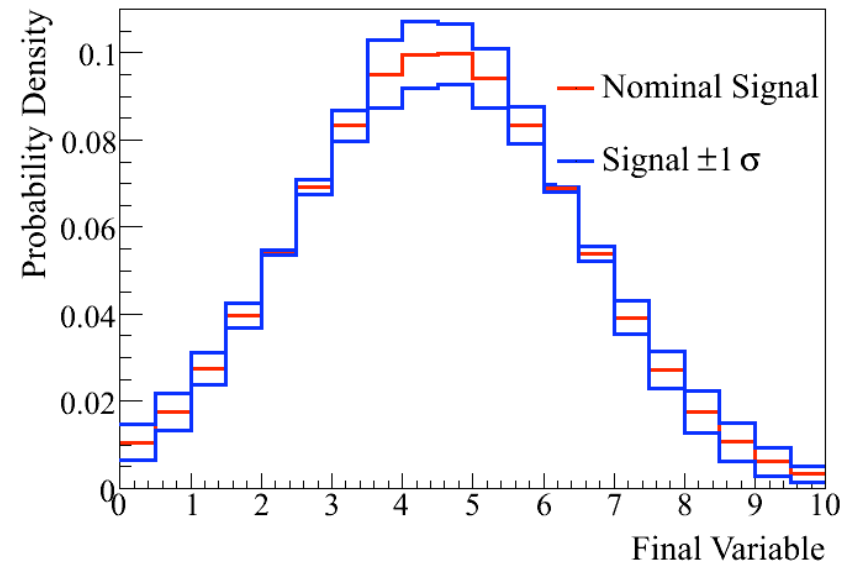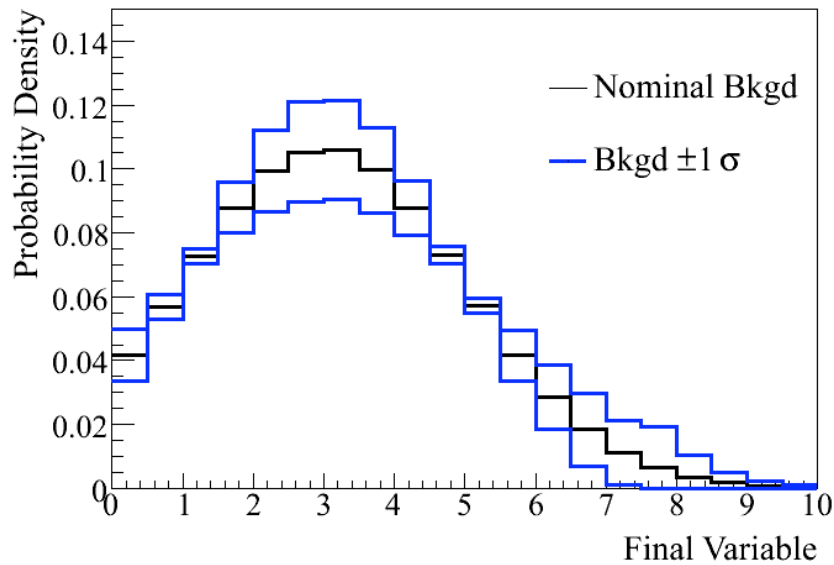
Challenge Aspects

The metric of the challenge:

We want to evaluate a measure of the signal significance for each problem. EG, a p-Value for the background-only hypothesis. Other metrics could be motivated.

Three aspects to the challenge:

A typical data analysis would proceed in three steps:

1) Ignore nuisance parameters as an analysis optimization step: "What is the best final variable configuration to find/measure this signal?"

2) Add nuisance parameters and determine your signal signficance.

3) Combine your results with another search (different analysis and/or another experiment).

What do we learn?

Which methods are most robust (eg, coverage for the parameter of interest)

Which methods are least sensitive to uncertainty on nuisance parameters

Which methods gain the most from additional signal-like search channels.

Etc.