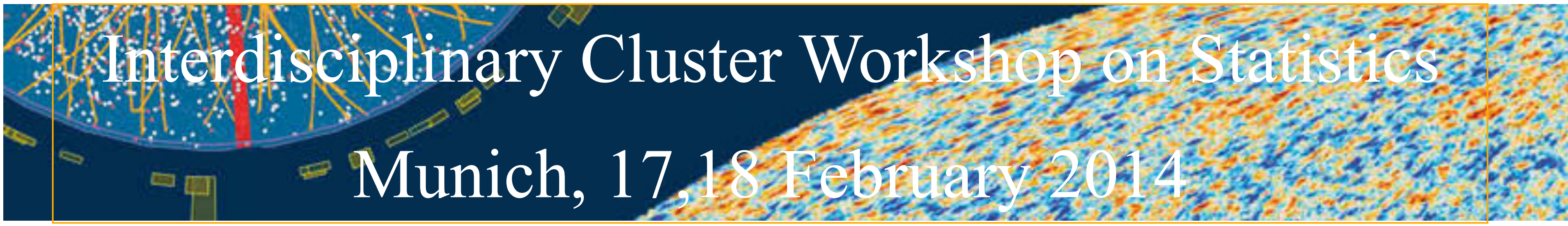
Overview of Statistical Methods in Particle Physics



http://intern.universe-cluster.de/indico/conferenceDisplay.py?confId=3022

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

I. Particle physics context

II. Tools

III. Multivariate methods

IV. Bayesian vs. Frequentist issues

discovery

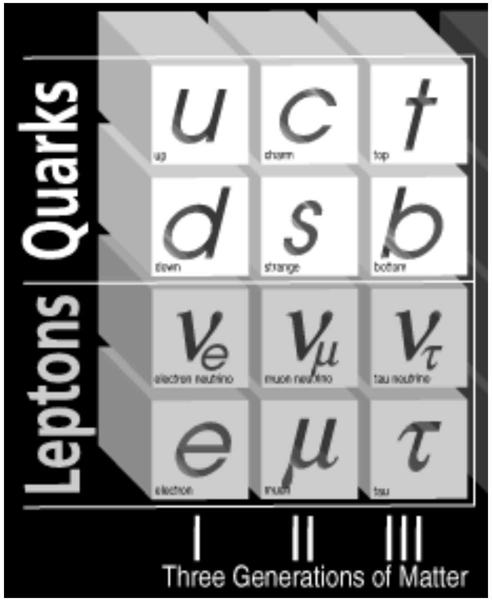
(exclusion limits)

V. Treatment of nuisance parameters

VI. Conclusions

The Standard Model of Particle Physics

Matter...



\+ force carriers...

photon (γ)
$W^{\pm}$
Z
gluon (g)

\+ Higgs boson

\+ relativity + quantum mechanics + gauge symmetries...

= “The Standard Model” = (usually) the null hypothesis H_0

Unanswered Questions

Why so many free parameters (~25)?

”Why vastly different energy scales? (naturalness, hierarchy,...)

CP violation & Baryon Asymmetry of Universe

Dark Matter? (Dark Energy???), etc.

No one believes SM is “true”, but ~everyone believes it is the correct “effective theory” valid at low energy.

Beyond the Standard Model (alternative hypothesis H_1):

Supersymmetry,

Extra gauge bosons (W’, *Z*’, ...),

Grand Unified Theories, Extra dimensions,

Compositeness, Micro black holes,...

Some particle physics experiments

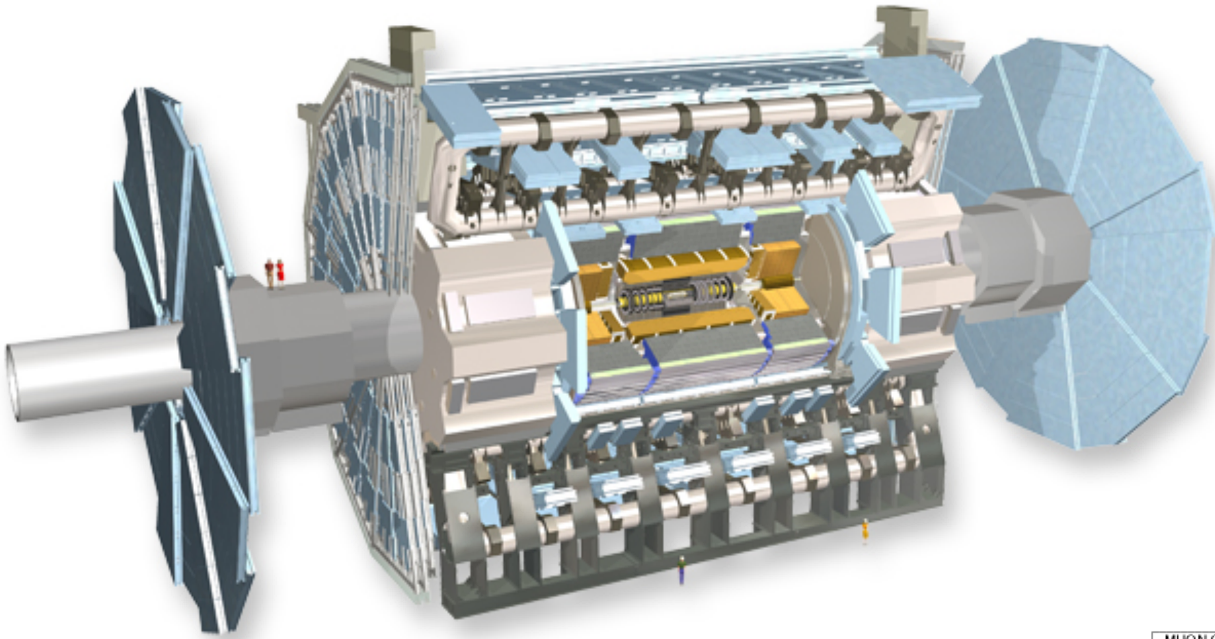
Large Hadron Collider: proton-proton collisions at CERN,
Neutrino physics (Fermilab, CERN, J-PARC),
Search for rare processes, e.g., double β-decay,
Searches for Dark Matter,
High-Energy Cosmic Rays,
Astrophysical Neutrinos,...

Earlier: proton-antiproton at Fermilab,
e^+e^- at SLAC, KEK, CERN (LEP),
e-p at DESY (HERA),...

Future: International Linear Collider (e^+e^-),
Future Circular Colliders (e^+e^-, pp),
High intensity neutrino beams,...

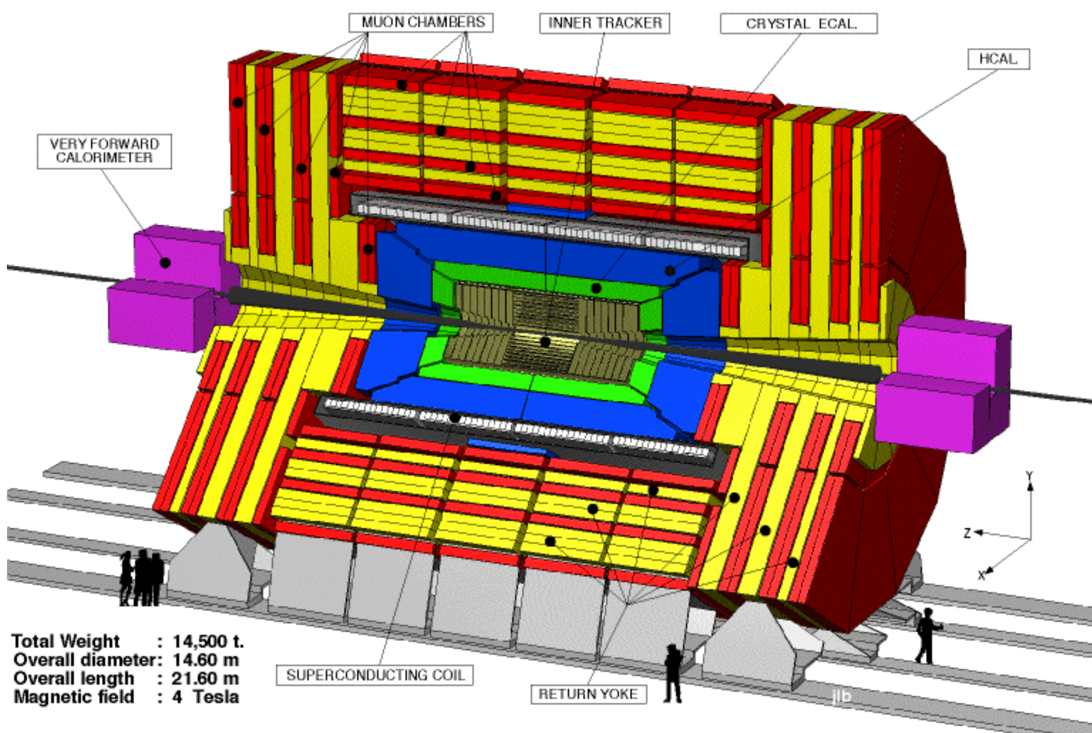
For today most examples use LHC physics for prototype analysis but methods usually applicable to other experiments as well.

The Large Hadron Collider at CERN



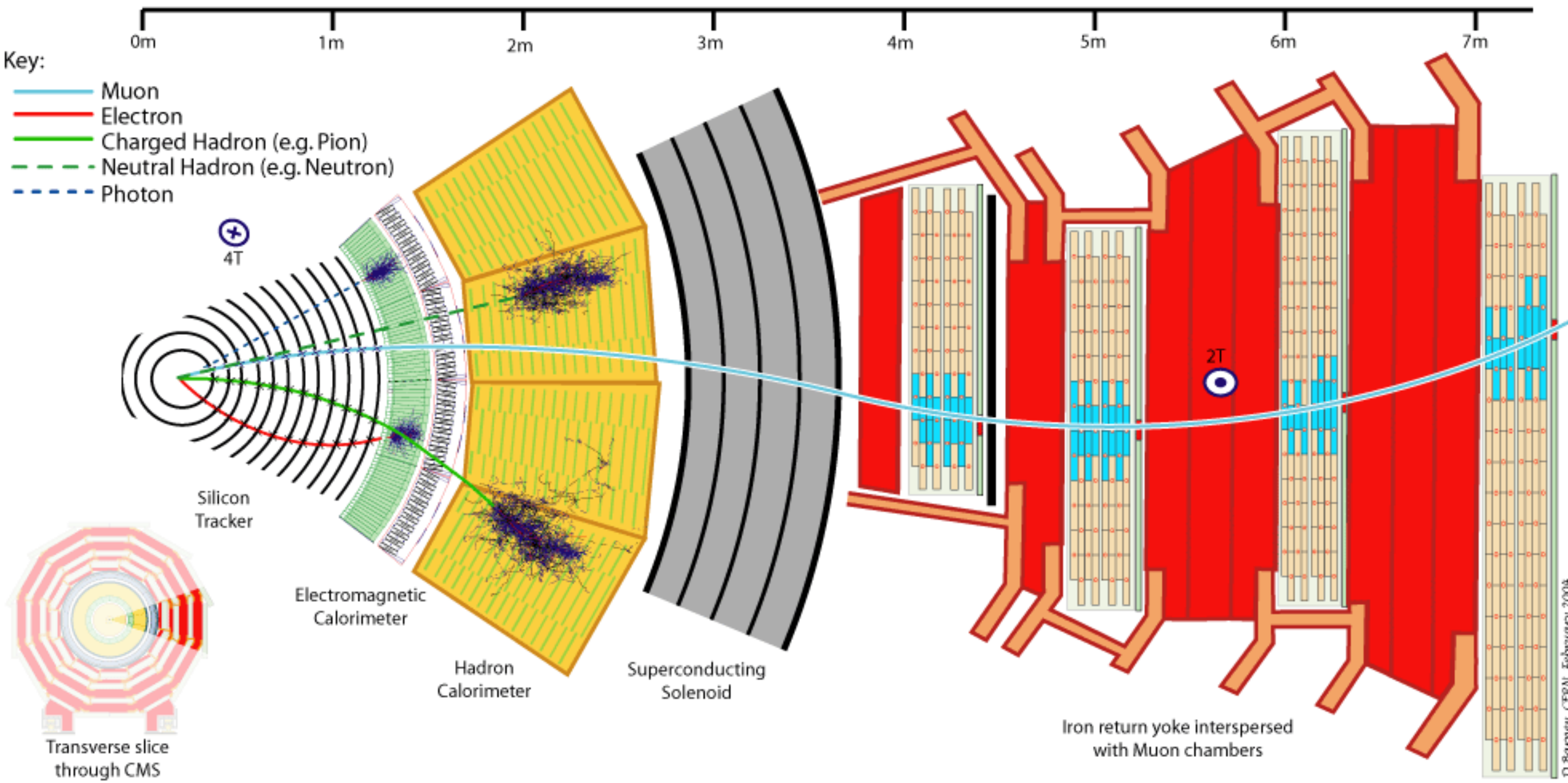
ATLAS

CMS

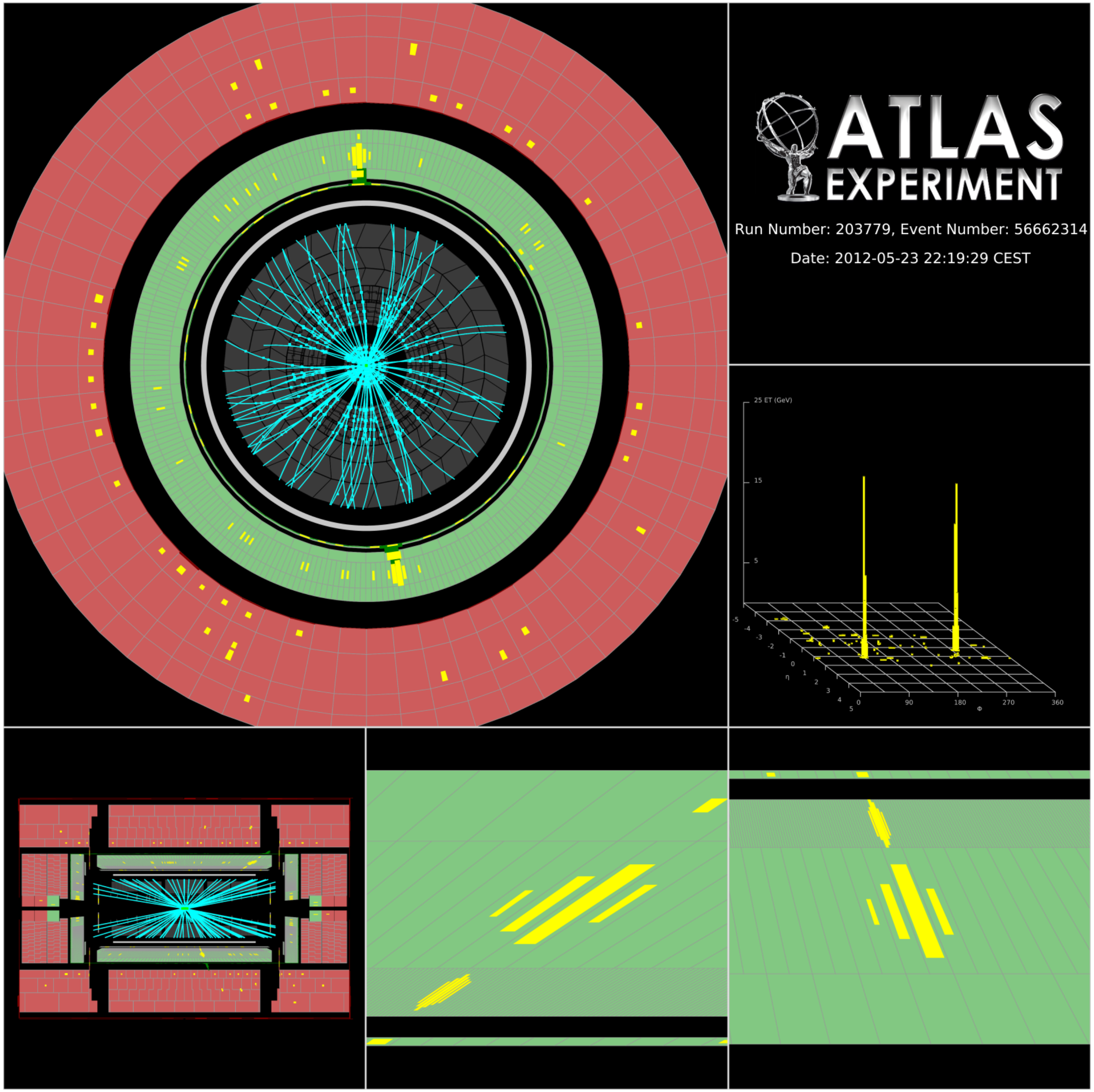


Also:
LHCb (B physics),
ALICE (heavy ions)

Detecting particles (CMS)

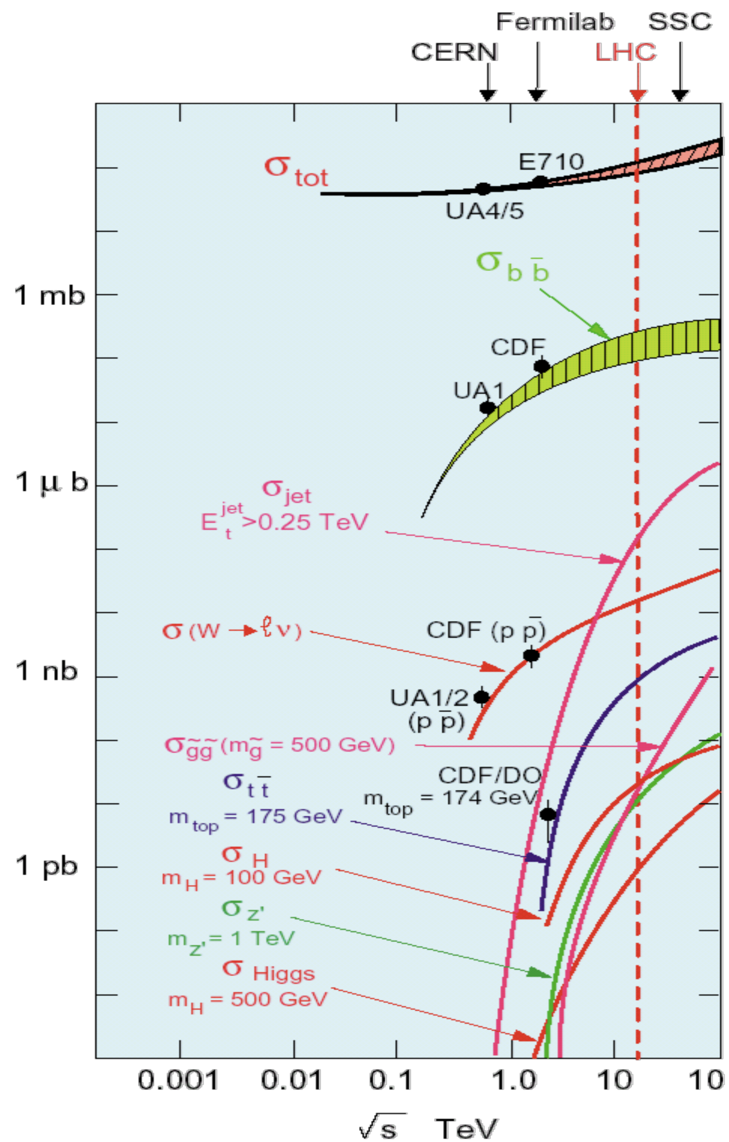


An "event"



Possible Higgs boson decay to two photons in ATLAS.

Event rates



Rate of "boring" events (background processes) ~GHz (after quick online sifting, record events at several × 10^2 Hz).

Rate of sought after signal process usually lower by many orders of magnitude.

Event size ~ Mbyte(s)

Calculating P(data | model)

Standard Model gives in principle probability of any data outcome.
In practice, many approximations needed:

perturbation theory (finite order + improvements),

parton densities (internal structure of colliding protons),

hadronization (conversion of quarks/gluons to hadrons),

detector simulation,...

Route to P(data|model) is usually through a Monte Carlo event generator combined with MC simulation program.

MC data → Training data for multivariate methods

Histogram MC → distributions of kinematic variables

Fundamentally, the analyses are based on *counting events*.

Usually new signal process *increases* expected rate.

A simulated event

⋮

```
                      Event listing (summary)

   I particle/jet KS     KF  orig    p_x      p_y      p_z       E        m

   1 !p+!         21    2212    0    0.000    0.000 7000.000 7000.000    0.938
   2 !p+!         21    2212    0    0.000    0.000-7000.000 7000.000    0.938
 ==============================================================================
   3 !g!          21      21    1    0.863   -0.323 1739.862 1739.862    0.000
   4 !ubar!       21      -2    2   -0.621   -0.163 -777.415  777.415    0.000
   5 !g!          21      21    3   -2.427    5.486 1487.857 1487.
   6 !g!          21      21    4  -62.910   63.357 -463.274  471.
   7 !~g!         21 1000021    0  314.363  544.843  498.897  979.
   8 !~g!         21 1000021    0 -379.700 -476.000  525.686  980.
   9 !~chi_1-!    21-1000024    7  130.058  112.247  129.860  263.
  10 !sbar!       21      -3    7  259.400  187.468   83.100  330.
  11 !c!          21       4    7  -79.403  242.409  283.026  381.
  12 !~chi_20!    21 1000023    8 -326.241  -80.971  113.712  385.
  13 !b!          21       5    8  -51.841 -294.077  389.853  491.
  14 !bbar!       21      -5    8   -0.597  -99.577   21.299  101.
  15 !~chi_10!    21 1000022    9  103.352   81.316   83.457  175.
  16 !s!          21       3    9    5.451   38.374   52.302   65.
  17 !cbar!       21      -4    9   20.839   -7.250   -5.938   22.
  18 !~chi_10!    21 1000022   12 -136.266  -72.961   53.246  181.
  19 !nu_mu!      21      14   12  -78.263  -24.757   21.719   84.
  20 !nu_mubar!   21     -14   12 -107.801   16.901   38.226  115.
 ===========================================================
  21 gamma         1      22    4    2.636    1.357    0.125    2.
  22 (~chi_1-)    11-1000024    9  129.643  112.440  129.820  262.
  23 (~chi_20)    11 1000023   12 -322.330  -80.817  113.191  382.
  24 ~chi_10       1 1000022   15   97.944   77.819   80.917  169.
  25 ~chi_10       1 1000022   18 -136.266  -72.961   53.246  181.
  26 nu_mu         1      14   19  -78.263  -24.757   21.719   84.
  27 nu_mubar      1     -14   20 -107.801   16.901   38.226  115.
  28 (Delta++)    11    2224    2    0.222    0.012-2734.287 2734.
:
```

```
 397 pi+            1      211  209    0.006    0.398 -308.296  308.297    0.140
 398 gamma          1       22  211    0.407    0.087-1695.458 1695.458    0.000
 399 gamma          1       22  211    0.113   -0.029 -314.822  314.822    0.000
 400 (pi0)         11      111  212    0.021    0.122 -103.709  103.709    0.135
 401 (pi0)         11      111  212    0.084   -0.068  -94.276   94.276    0.135
 402 (pi0)         11      111  212    0.267   -0.052 -144.673  144.674    0.135
 403 gamma          1       22  215   -1.581    2.473    3.306    4.421    0.000
 404 gamma          1       22  215   -1.494    2.143    3.051    4.016    0.000
 405 pi-            1     -211  216    0.007    0.738    4.015    4.085    0.140
 406 pi+            1      211  216   -0.024    0.293    0.486    0.585    0.140
 407 K+             1      321  218    4.382   -1.412   -1.799    4.968    0.494
 408 pi-            1     -211  218    1.183   -0.894   -0.176    1.500    0.140
 409 (pi0)         11      111  218    0.955   -0.459   -0.590    1.221    0.135
 410 (pi0)         11      111  218    2.349   -1.105   -1.181    2.855    0.135
 411 (Kbar0)       11     -311  219    1.441   -0.247   -0.472    1.615    0.498
 412 pi-            1     -211  219    2.232   -0.400   -0.249    2.285    0.140
 413 K+             1      321  220    1.380   -0.652   -0.361    1.644    0.494
 414 (pi0)         11      111  220    1.078   -0.265    0.175    1.132    0.135
 415 (K_S0)        11      310  222    1.841    0.111    0.894    2.109    0.498
 416 K+             1      321  223    0.307    0.107    0.252    0.642    0.494
 417 pi-            1     -211  223    0.266    0.316   -0.201    0.480    0.140
 418 nbar0          1    -2112  226    1.335    1.641    2.078    3.111    0.940
 419 (pi0)         11      111  226    0.899    1.046    1.311    1.908    0.135
 420 pi+            1      211  227    0.217    1.407    1.356    1.971    0.140
 421 (pi0)         11      111  227    1.207    2.336    2.767    3.820    0.135
 422 n0             1     2112  228    3.475    5.324    5.702    8.592    0.940
 423 pi-            1     -211  228    1.856    2.606    2.808    4.259    0.140
 424 gamma          1       22  229   -0.012    0.247    0.421    0.489    0.000
 425 gamma          1       22  229    0.025    0.034    0.009    0.043    0.000
 426 pi+            1      211  230    2.718    5.229    6.403    8.703    0.140
 427 (pi0)         11      111  230    4.109    6.747    7.597   10.961    0.135
 428 pi-            1     -211  231    0.551    1.233    1.945    2.372    0.140
 429 (pi0)         11      111  231    0.645    1.141    0.922    1.608    0.135
 430 gamma          1       22  232   -0.383    1.169    1.208    1.724    0.000
 431 gamma          1       22  232   -0.201    0.070    0.060    0.221    0.000
:
```

⋮

PYTHIA Monte Carlo
pp → gluino-gluino

Monte Carlo detector simulation

Takes as input the particle list and momenta from generator.

Simulates detector response:

multiple Coulomb scattering (generate scattering angle),
particle decays (generate lifetime),
ionization energy loss (generate Δ),
electromagnetic, hadronic showers,
production of signals, electronics response, ...

Output = simulated raw events $\rightarrow$ input to reconstruction software:

track finding, fitting, etc.

Programming package: **`GEANT`**

Events usually have weight, e.g., to correct for effects not known at the time of initial simulation.

Prototypical analyses

Select events with properties characteristic of signal process (invariably select some background events as well).

Case #1:

Existence of signal process already well established (e.g. production of top quarks)

Study properties of signal events (e.g., measure top quark mass, production cross section, decay properties,...)

Statistics issues:

Event selection → multivariate classifiers
Parameter estimation
(usually maximum likelihood or least squares)
Bias, variance of estimators; goodness-of-fit
Unfolding (deconvolution).

Prototypical analyses (cont.): a "search"

Case #2:

Existence of signal process not yet established.

Goal is to see if it exists by rejecting the background-only hypothesis.

H_0: All of the selected events are background (usually means "standard model" or events from known processes)

H_1: Selected events contain a mixture of background and signal.

Statistics issues:

Optimality (power) of statistical test.
Rejection of H_0 usually based on p-value $< 2.9 \times 10^{-7}$ (5σ).
Some recent interest in use of Bayes factors.

In absence of discovery, exclusion limits on parameters of signal models (frequentist, Bayesian, "CLs",...)

Tools

C++ (plus some python, matlab,...)

Transition from FORTRAN late 1990s

ROOT

`root.cern.ch`

Data analysis program & C++ library

Developed at CERN late 1990s (Brun, Rademakers,...)

Many utilities for manipulation of *n*-tuples, histograms, parameter fitting (Minuit), numerical methods, visualization

Program commands are interpreted C++

Workhorse analysis environment in particle physics; used as basis for RooFit, RooStats

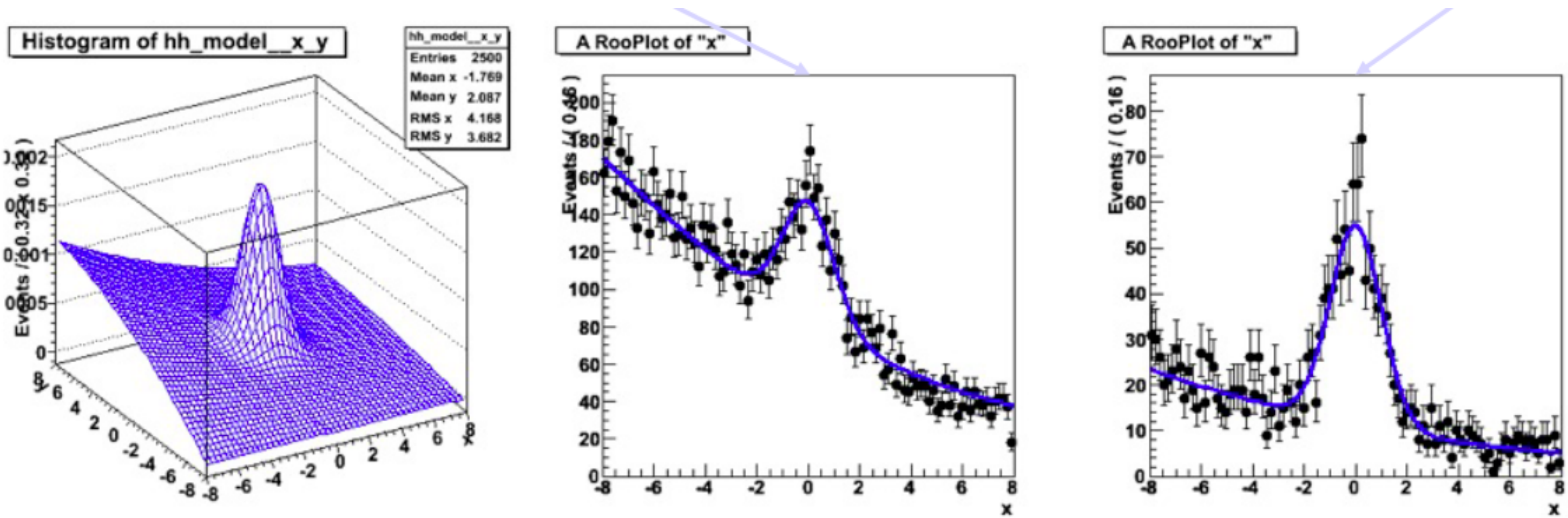
RooFit

Developed in BaBar experiment (W. Verkerke, D. Kirkby)

ROOT-based toolkit for fitting and building statistical models plus utilities for ML, LS fitting.

Utilities for pdf normalization, convolutions

Contains “workspaces” that can store the data plus underlying statistical model in a root file.



RooStats

C++ library for statistical analyses, arXiv:1009.1003

ATLAS/CMS collaborative effort to consolidate statistical tools for LHC analyses

Cranmer, Verkerke, Schott, Moneta, + many developers

Uses ROOT, RooFit (especially workspaces)

Tools for combination of measurements, e.g., searches for the HIggs that combine several decay modes.

Implements a wide variety of statistical methods

Frequentist hypothesis tests and confidence intervals

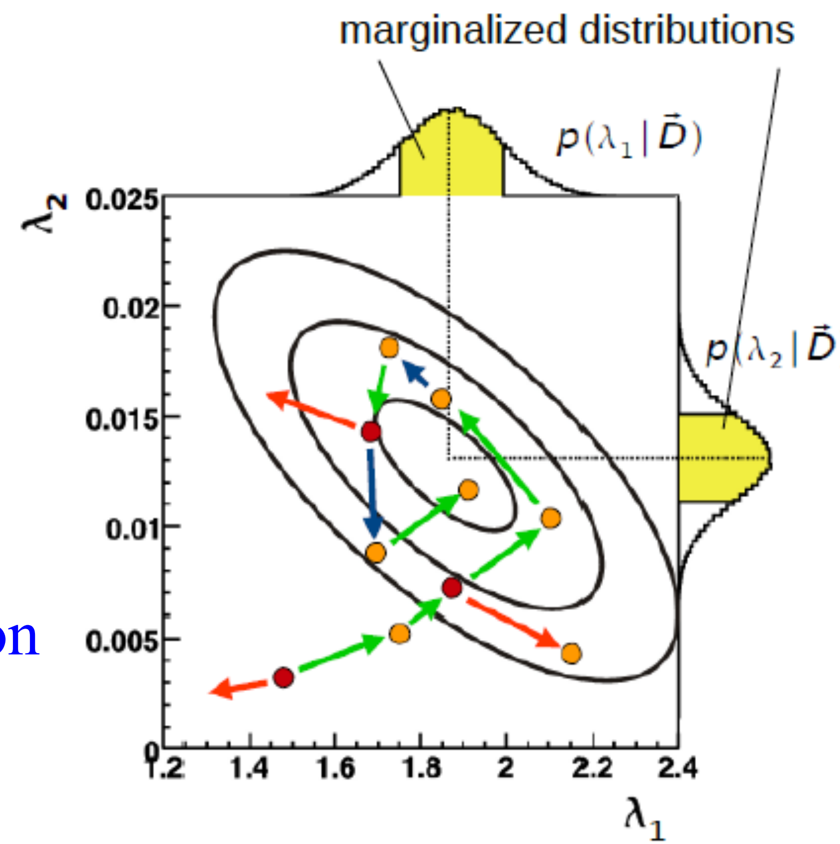
Bayesian analyses, MCMC (interface to BAT)

Bayesian Analysis Toolkit

A. Caldwell, D. Kollar, K. Kröninger, Comp. Phys. Comm. 180 (2009) 2197-2209; arXiv:0808.2552

General framework specifically for Bayesian computation, especially MCMC for marginalizing posterior probabilities.

Support for Bayes factors through interface to "Cuba" library for multidimensional numerical integration (T. Hahn, Comp.Phys.Comm. 168 (2005) 78-95, arXiv:hep-ph/0404043)



Multivariate methods: intro

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons,

x_2 = mean p_T of jets,

x_3 = missing energy, ...

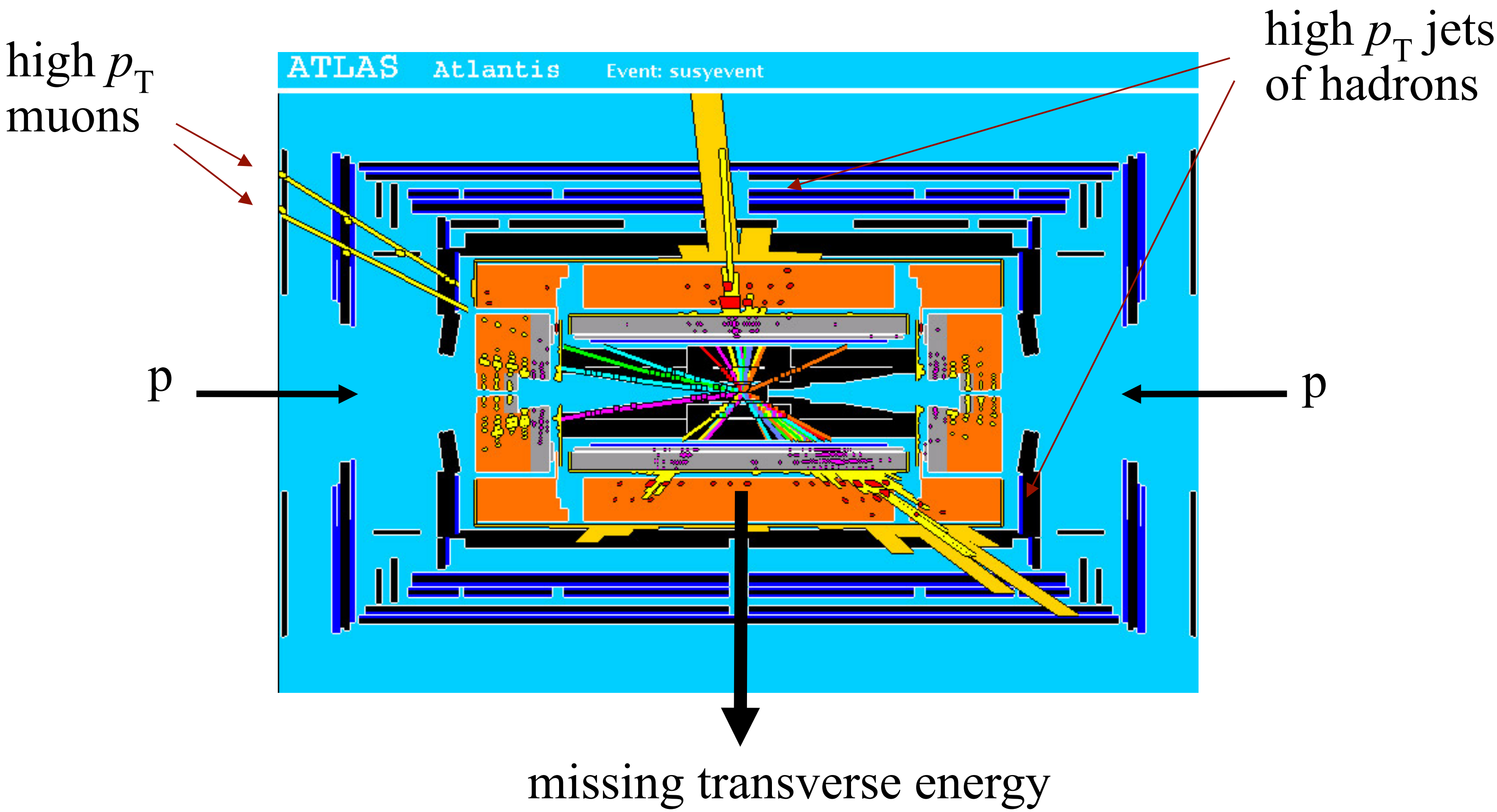
$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$\mathrm{pp} \rightarrow t\bar{t}\ , \quad \mathrm{pp} \rightarrow \tilde{g}\tilde{g}\ , \ldots$$

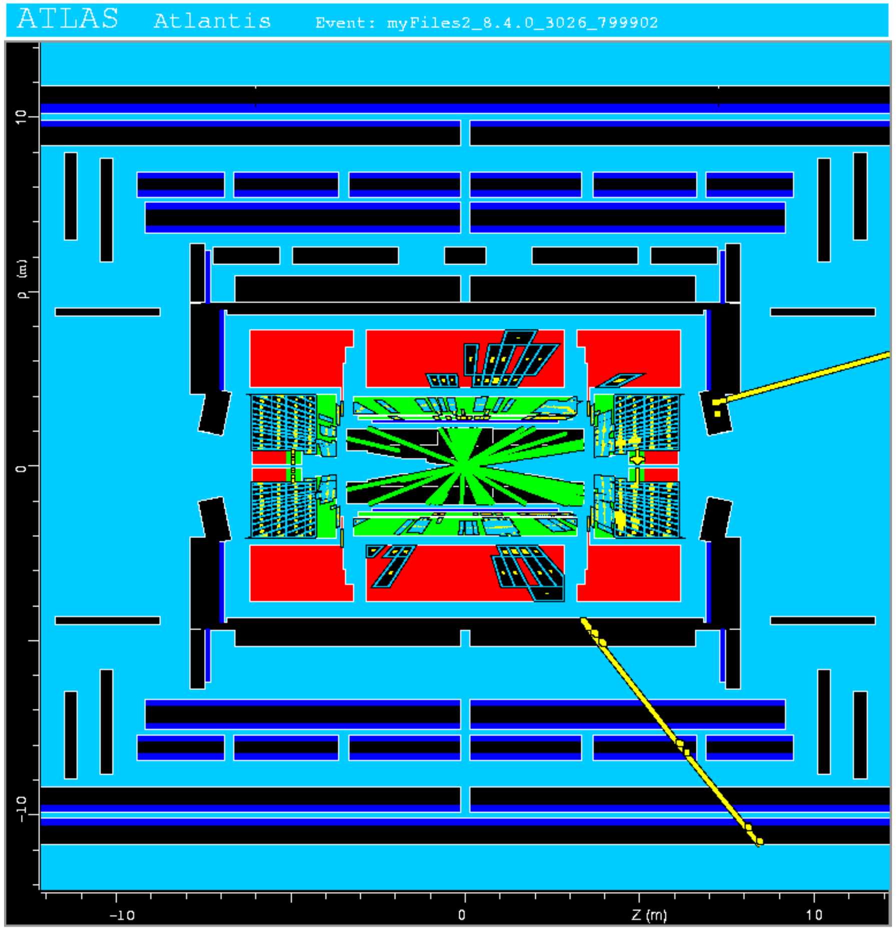
For each reaction we consider we will have a hypothesis for the pdf of $\vec{x}$, e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

E.g. call H_0 the background hypothesis (the event type we want to reject); H_1 is signal hypothesis (the type we want).

A simulated SUSY event in ATLAS



Background events



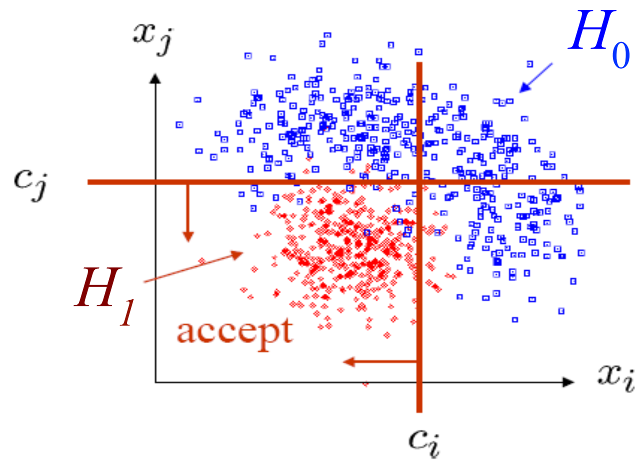
This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

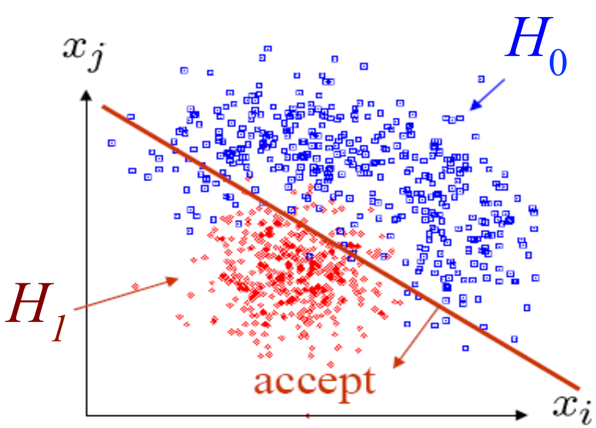
Defining critical region of test

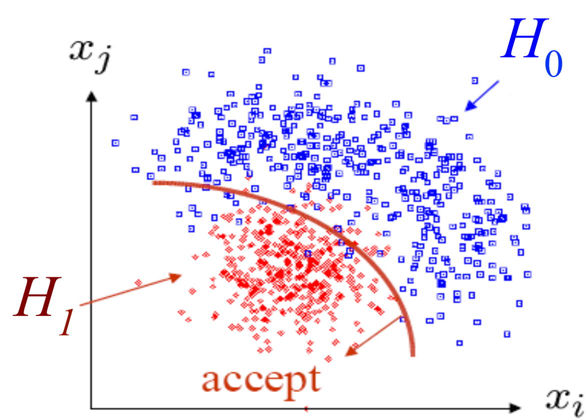
In particle physics usually start by making simple "cuts":

$$x_i < c_i$$
$$x_j < c_j$$



Maybe later try some other type of decision boundary:





Multivariate methods

Many new (and some old) methods:

- Fisher discriminant
- Neural networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

Software used in HEP:

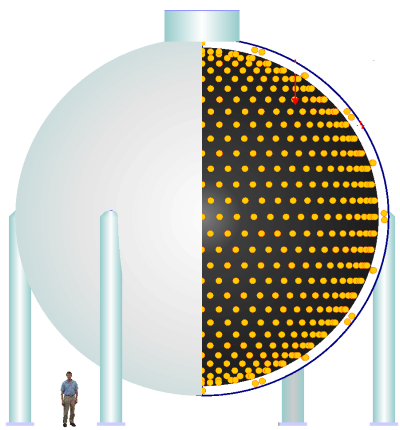
TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

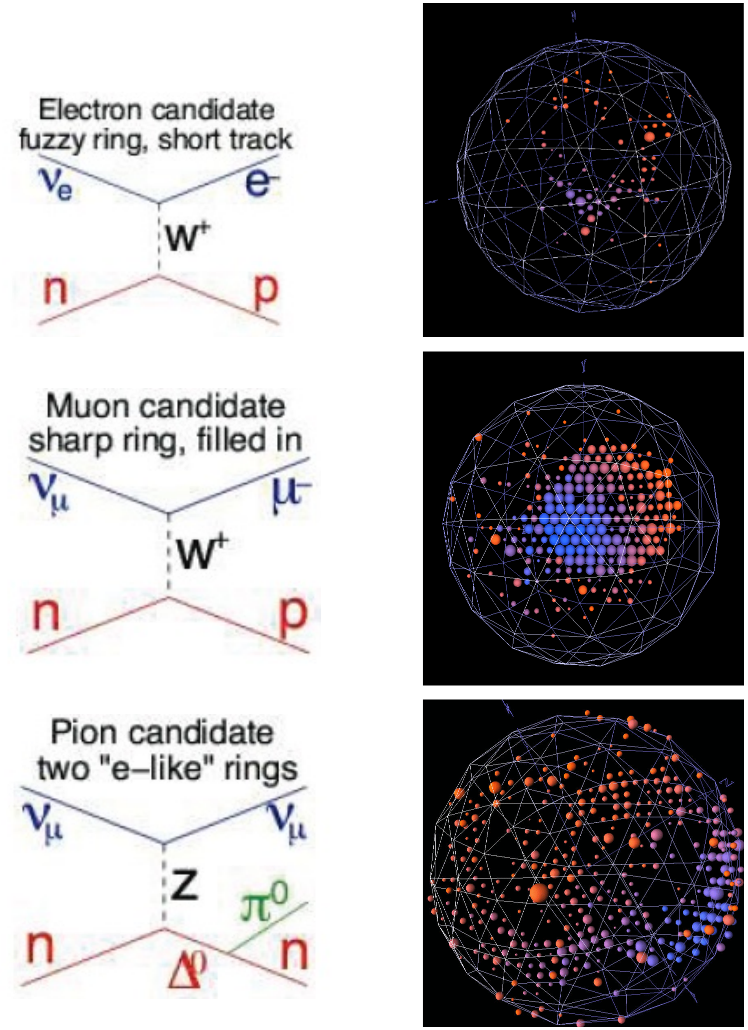
StatPatternRecognition, I. Narsky, physics/0507143

BDT example: particle i.d. in MiniBooNE

Detector is a 12-m diameter tank of mineral oil exposed to a beam of neutrinos and viewed by 1520 photomultiplier tubes:

MiniBooNE Detector





H.J. Yang, MiniBooNE PID, DNP06

Search for ν_μ to ν_e oscillations required particle i.d. using information from the PMTs.

Decision trees

Start with Monte Carlo training data; event types (signal/background known.

Out of all the input variables, find the one for which with a single cut gives best improvement in signal purity:

$$P = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}$$

where w_i. is the weight of the ith event.

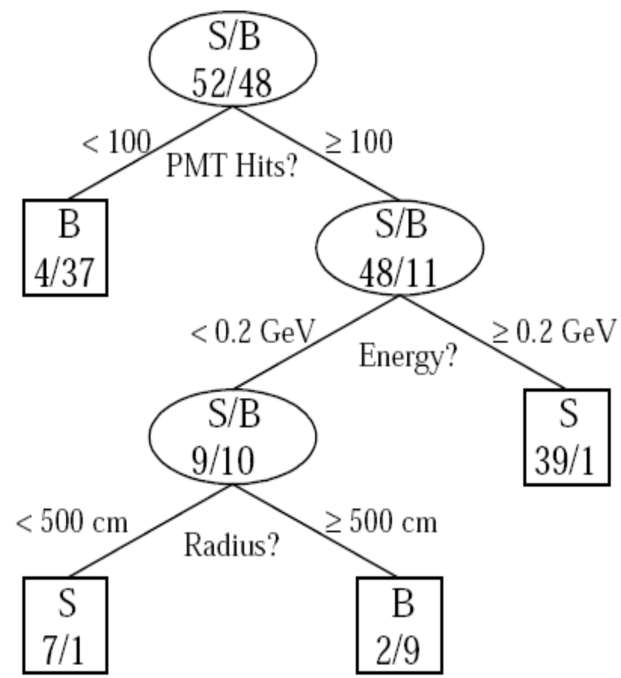
Resulting nodes classified as either signal/background.

Iterate until stop criterion reached based on e.g. purity or minimum number of events in a node.

The set of cuts defines classifier; let

$f(\boldsymbol{x}) = 1$ on signal side,

$f(\boldsymbol{x}) = -1$ on background side.



Example by MiniBooNE experiment, B. Roe et al., NIM 543 (2005) 577

Boosting

General method for creating set of classifiers that are averaged.

Classifier must admit event weights (e.g., decision tree).

After training a classifier, update the weights according to a given rule so as to increase the weights of the incorrectly classified events.

$$w_i^{(k+1)} = w_i^{(k)} \frac{e^{-\alpha_k f_k(\boldsymbol{x}_i) y_i / 2}}{Z_k}$$

Widely studied example: AdaBoost (Freund and Shapire, 1997): At *k*th iteration, assign score

$$\alpha_k = \ln \frac{1-\varepsilon_k}{\varepsilon_k}$$

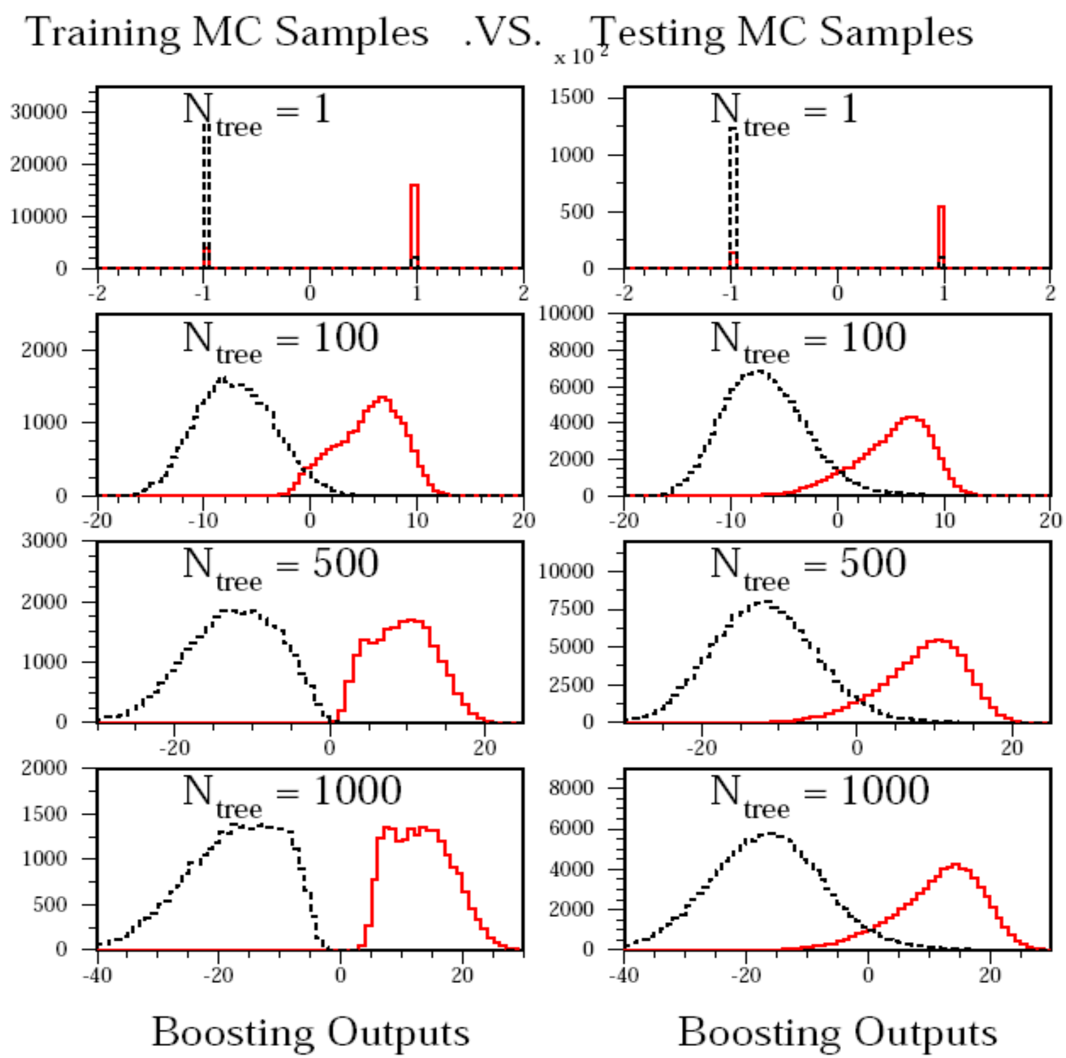
misclassification rate

Final classifier using K iterations: $$f(\boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k f_k(\boldsymbol{x}, T_k)$$

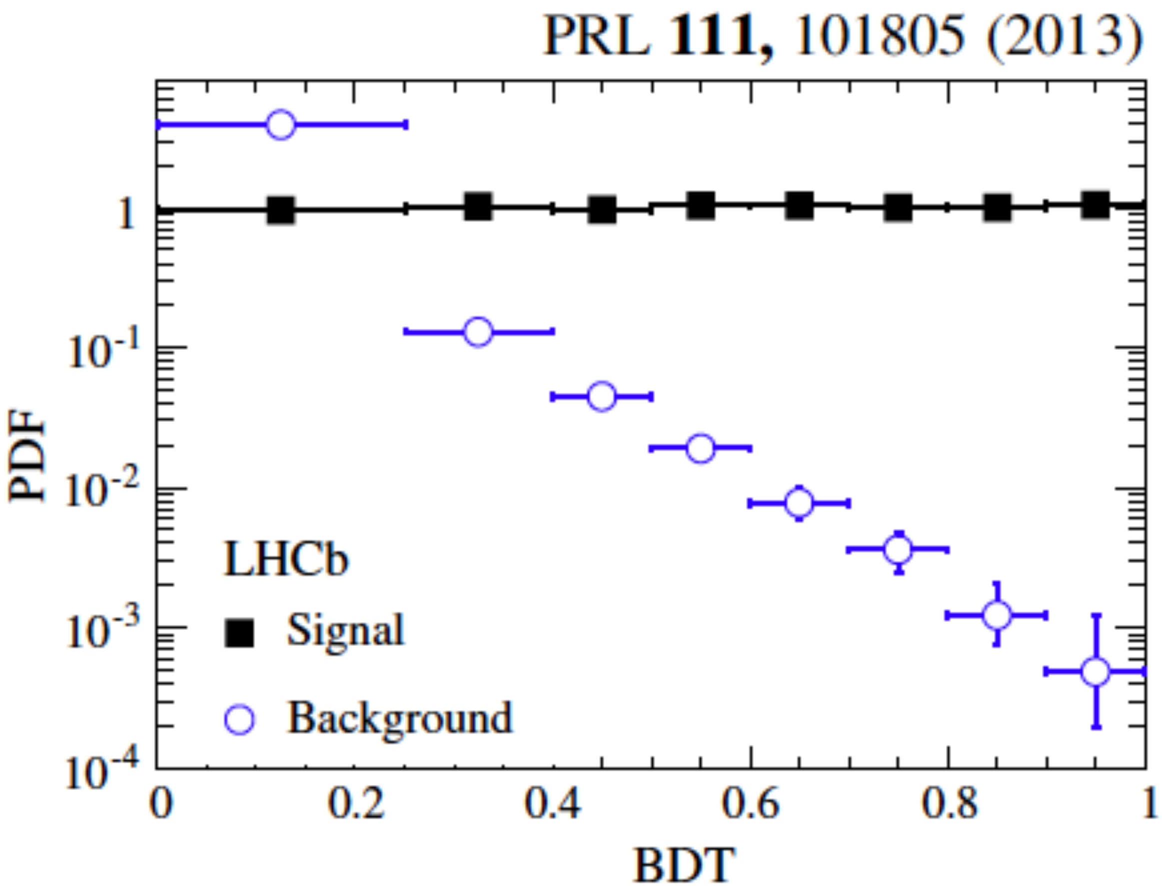
Monitoring overtraining

From MiniBooNE example:

Performance stable after a few hundred trees.



BDT Example: observation of $B_s \rightarrow \mu\mu$ by the LHCb Experiment



Signal rate determined by fits to mass dist. within each BDT bin.

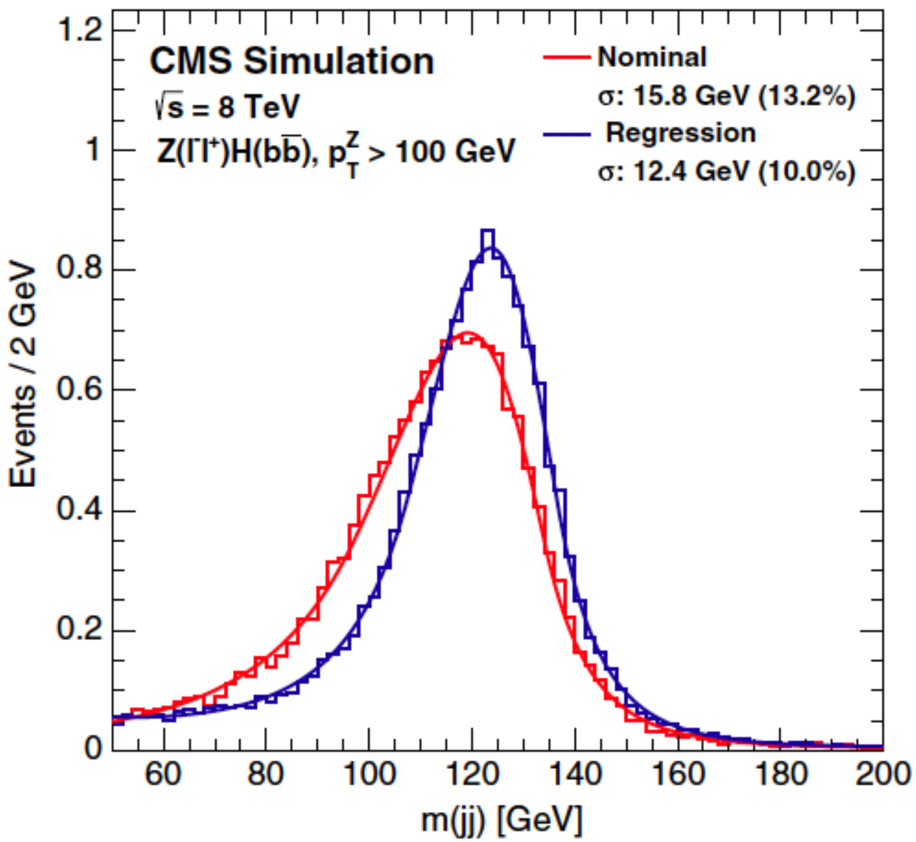
$$\mathcal{B}(B_s^0 \to \mu^+\mu^-) = (2.9^{+1.1}_{-1.0}) \times 10^{-9}$$

Multivariate regression

Example: a jet of hadrons initiated by a b quark deposits particles of different types and energies in the detector than one caused by a gluon or light quark (e.g., sometimes b-jets contain unseen neutrinos together with charged leptons).

Feed several measured jet properties into neural network function to estimate energy. (Aaltonen et al., arXiv:1107.3026)

Used by CMS search for H→bb: PRD, 89, 012003 (2014)



Bayesian vs. Frequentist Approaches

Frequentist:

Probability only assigned to (repeatable) data outcomes.

Results depend on probability for data found and not found, e.g.,

p-value = P(data equal or more extreme | H)

Discovery (roughly): p-value of no signal hypothesis < α.

Bayesian:

Use subjective probability for P(model|data).

Obtain from P(data|model) with Bayes' theorem;

needs prior probability P(model).

Discovery can be based on Bayes factor:

$B_{01} = P(\text{data}|H_0) \,/\, P(\text{data}|H_1)$ (requires only prior for model's internal parameters).

Traditionally most Particle Physics analyses use frequentist approach; some use of Bayesian limits and Bayes factors.

Prototype frequentist search analysis

Search for signal in a region of phase space; result is histogram of some variable *x* giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx \,, \qquad b_i = b_{\rm tot} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_{\mathrm{s}}$, $\boldsymbol{\theta}_{\mathrm{b}}$,$b_{\mathrm{tot}}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

maximizes L for
Specified μ

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximize L

The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

The profile LR hould be near-optimal in present analysis with variable μ and nuisance parameters $\boldsymbol{\theta}$.

Test statistic for discovery

Try to reject background-only ($\mu = 0$) hypothesis using

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only regard upward fluctuation of data as evidence against the background-only hypothesis.

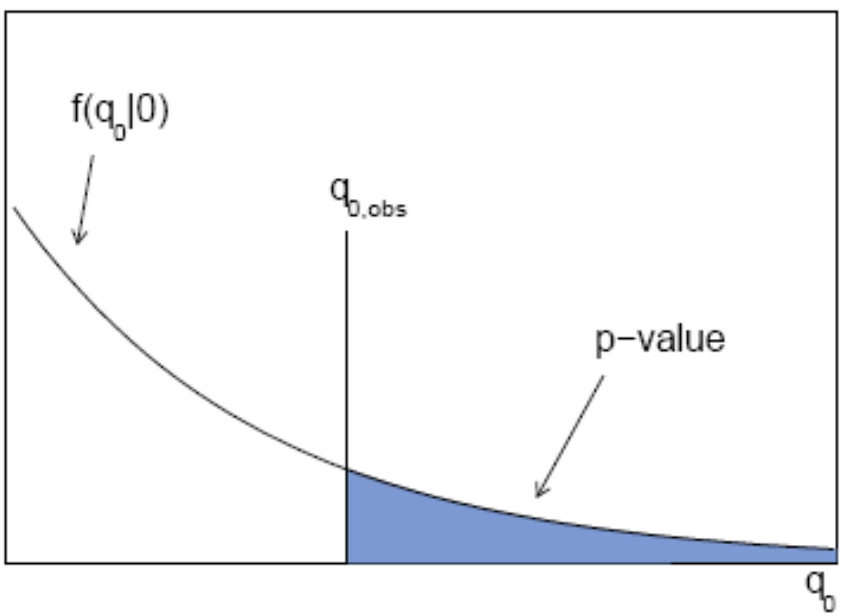
Note that even if physical models have $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\mathrm{obs}}$ is

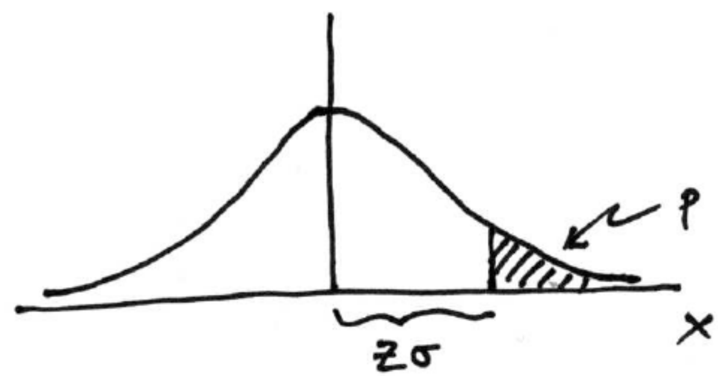
$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

need formula for this



From p-value get equivalent significance,

$$Z = \Phi^{-1}(1-p)$$



Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The p-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$\boxed{Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}}$$

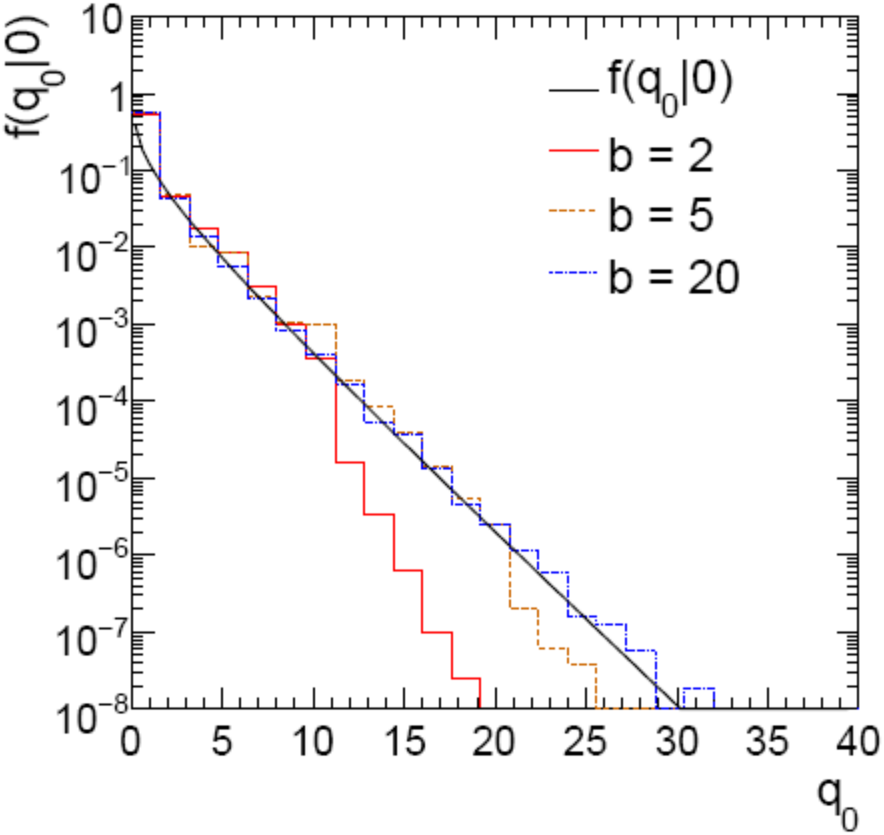
Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

$$m \sim \text{Poisson}(\tau b)$$

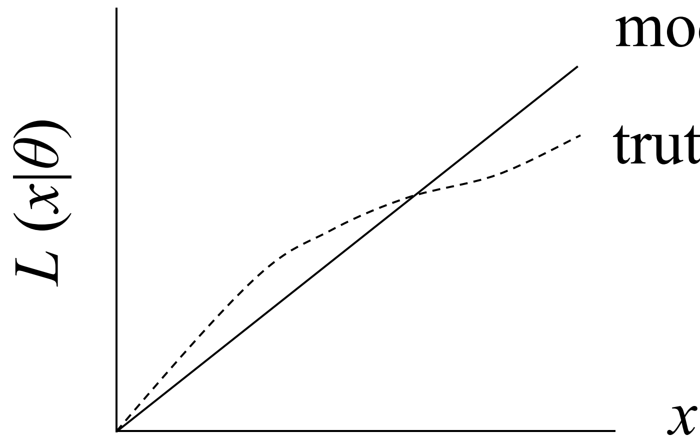
Here take $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.



Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \rightarrow L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

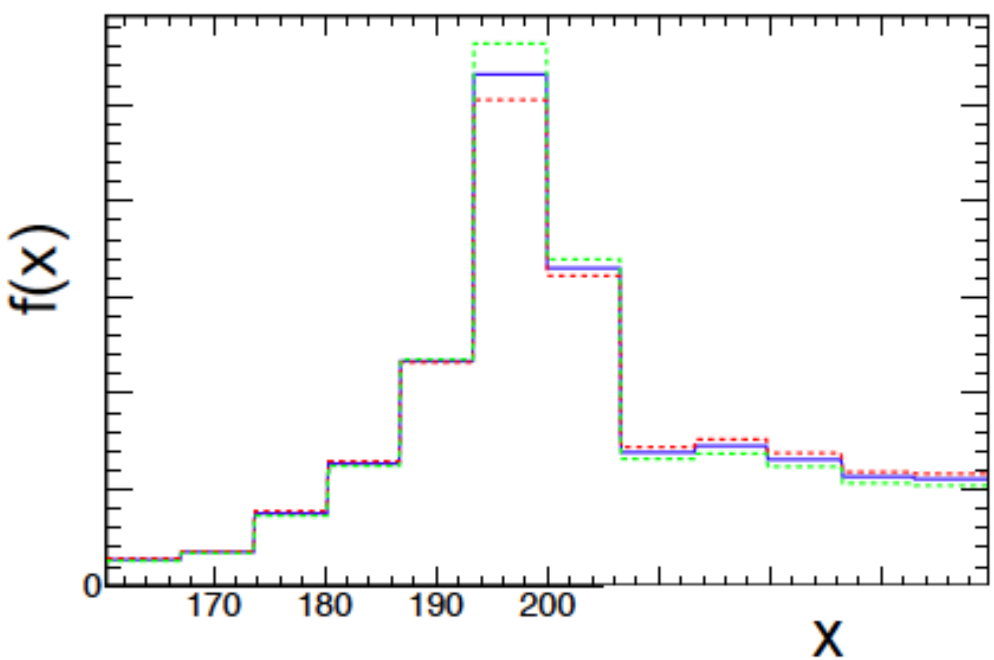
Example tool: HistFactory

Tool for creating statistical models for use with RooFit and RooStats

K. Cranmer, G. Lewis, L. Moneta, A. Shibata, W. Verkerke, CERN-OPEN-2012-016 01/01/2012

Often prediction for a distribution comes from MC, run with nominal parameter values.

K. Cranmer



Can also produce same distribution with alternative parameters (e.g., jet energy calibration constants, ...), typically vary "$\pm 1\sigma$".

Interpolate histograms to give parametric model containing new nuisance parameter.

Interpolating histograms

Interpolate histograms to give parametric model containing new nuisance parameter α, e.g.,

$\alpha = 0$:	nominal histogram
$\alpha = \pm 1$:	$\pm 1\sigma$ histogram

Treat nominal value of parameter α_0 (here zero) as a measurement,

$$P(x) \to P(x, \alpha_0|\alpha) = P(x|\alpha)\frac{1}{\sqrt{2\pi}\sigma}e^{-(\alpha_0-\alpha)^2/2\sigma^2}$$

Often use normal, or log-normal ($\lambda = \ln \alpha$, $\lambda \sim$ Gauss).

Or in Bayesian analysis, assign corresponding normal or log-normal prior $\pi(\alpha)$.

Related technology used earlier at Tevatron (MCLimit, Collie)

p-values in cases with nuisance parameters

Suppose we have a statistic q_θ that we use to test a hypothesized value of a parameter θ, such that the p-value of θ is

$$p_\theta = \int_{q_{\theta,\mathrm{obs}}}^{\infty} f(q_\theta|\theta,\nu)\,dq_\theta$$

But what values of ν to use for $f(q_\theta|\theta, \nu)$?

Fundamentally we want to reject θ only if $p_\theta < \alpha$ for all ν.

→ "exact" confidence interval

Recall that for statistics based on the profile likelihood ratio, the distribution $f(q_\theta|\theta, \nu)$ becomes independent of the nuisance parameters in the large-sample limit.

But in general for finite data samples this is not true; one may be unable to reject some θ values if all values of ν must be considered, even those strongly disfavoured by the data (resulting interval for θ "overcovers").

Profile construction ("hybrid resampling")

K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Approximate procedure is to reject θ if $p_\theta \leq \alpha$ where the p-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ:

$$\hat{\hat{\nu}}(\theta)$$

"double hat" notation means value of parameter that maximizes likelihood for the given θ.

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{\nu}}(\theta))$.

Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

“Hybrid frequentist-Bayesian” method

Alternatively, suppose uncertainty in ν is characterized by a Bayesian prior $\pi(\nu)$.

Can use the marginal likelihood to model the data:

$$L_{\rm m}(x|\theta) = \int L(x|\theta,\nu)\pi(\nu)\,d\nu$$

This does not represent what the data distribution would be if we “really” repeated the experiment, since then ν would not change.

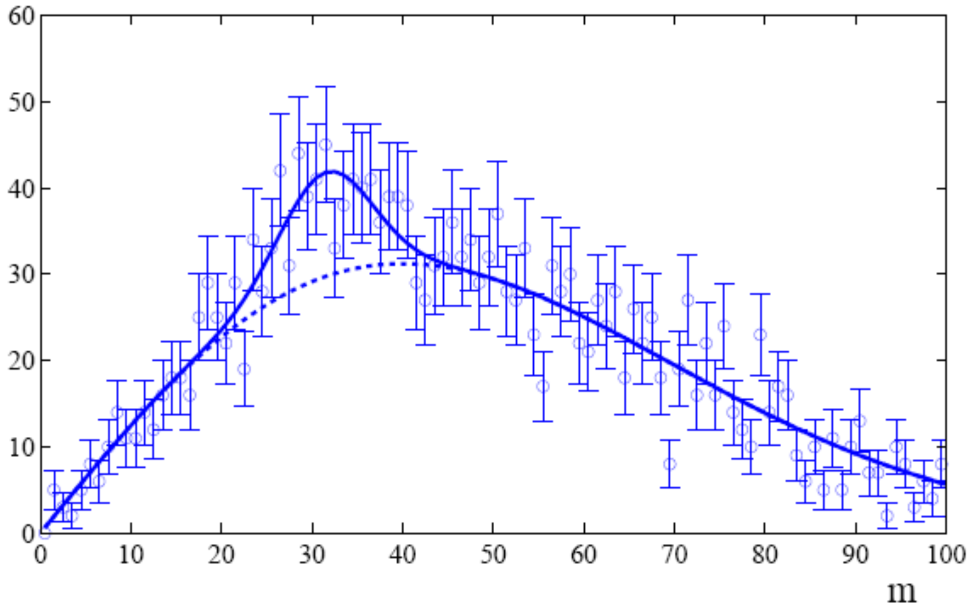
But the procedure has the desired effect. The marginal likelihood effectively builds the uncertainty due to ν into the model.

Use this now to compute (frequentist) p-values → the model being tested is in effect a weighted average of models.

The Look-Elsewhere Effect

Suppose a model for a mass distribution allows for a peak at a mass m with amplitude μ.

The data show a bump at a mass m_0.



How consistent is this with the no-bump ($\mu = 0$) hypothesis?

Local p-value

First, suppose the mass m_0 of the peak was specified a priori.

Test consistency of bump with the no-signal ($\mu = 0$) hypothesis with e.g. likelihood ratio

$$t_{\rm fix} = -2 \ln \frac{L(0, m_0)}{L(\hat{\mu}, m_0)}$$

where "fix" indicates that the mass of the peak is fixed to m_0.

The resulting p-value

$$p_{\rm local} = \int_{t_{\rm fix,obs}}^{\infty} f(t_{\rm fix}|0) \, dt_{\rm fix}$$

gives the probability to find a value of $t_{\rm fix}$ at least as great as observed at the specific mass m_0 and is called the local p-value.

Global *p*-value

But suppose we did not know where in the distribution to expect a peak.

What we want is the probability to find a peak at least as significant as the one observed anywhere in the distribution.

Include the mass as an adjustable parameter in the fit, test significance of peak using

$$t_{\text{float}} = -2\ln\frac{L(0)}{L(\hat{\mu},\hat{m})}$$

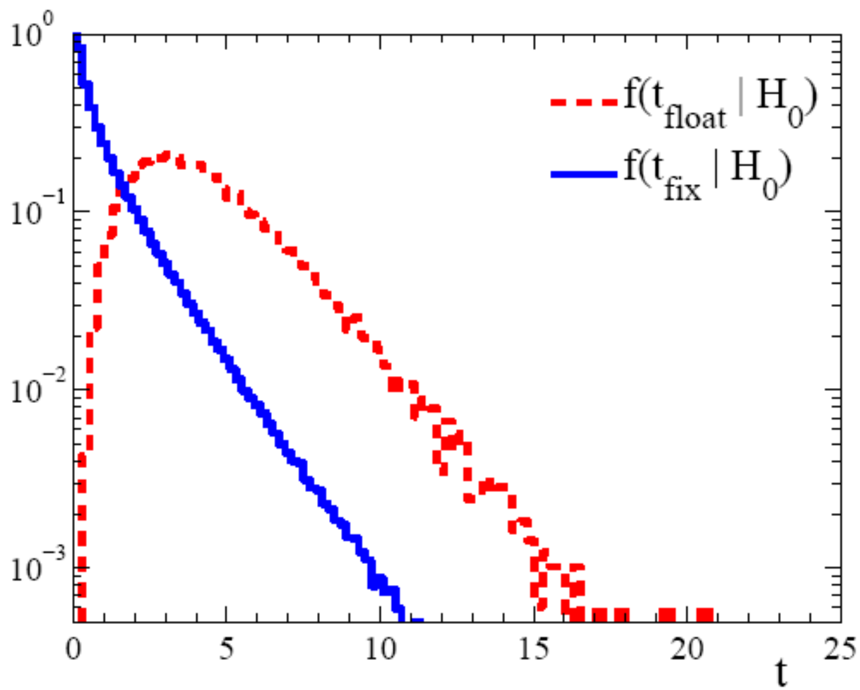
(Note m does not appear in the $\mu = 0$ model.)

$$p_{\text{global}} = \int_{t_{\text{float,obs}}}^{\infty} f(t_{\text{float}}|0)\,dt_{\text{float}}$$

Distributions of t_{fix}, t_{float}

For a sufficiently large data sample, t_{fix} ~chi-square for 1 degree of freedom (Wilks' theorem).

For t_{float} there are two adjustable parameters, μ and m, and naively Wilks theorem says t_{float} ~ chi-square for 2 d.o.f.



In fact Wilks' theorem does not hold in the floating mass case because on of the parameters (m) is not-defined in the $\mu = 0$ model.

So getting t_{float} distribution is more difficult.

Approximate correction for LEE

We would like to be able to relate the p-values for the fixed and floating mass analyses (at least approximately).

Gross and Vitells show the p-values are approximately related by

$$p_{\text{global}} \approx p_{\text{local}} + \langle N(c) \rangle$$

where $\langle N(c) \rangle$ is the mean number "upcrossings" of $t_{\text{fix}} = -2\ln\lambda$ in the fit range based on a threshold

$$c = t_{\text{fix,obs}} = Z_{\text{local}}^2$$

and where $Z_{\text{local}} = \Phi^{-1}(1 - p_{\text{local}})$ is the local significance.

So we can either carry out the full floating-mass analysis (e.g. use MC to get p-value), or do fixed mass analysis and apply a correction factor (much faster than MC).

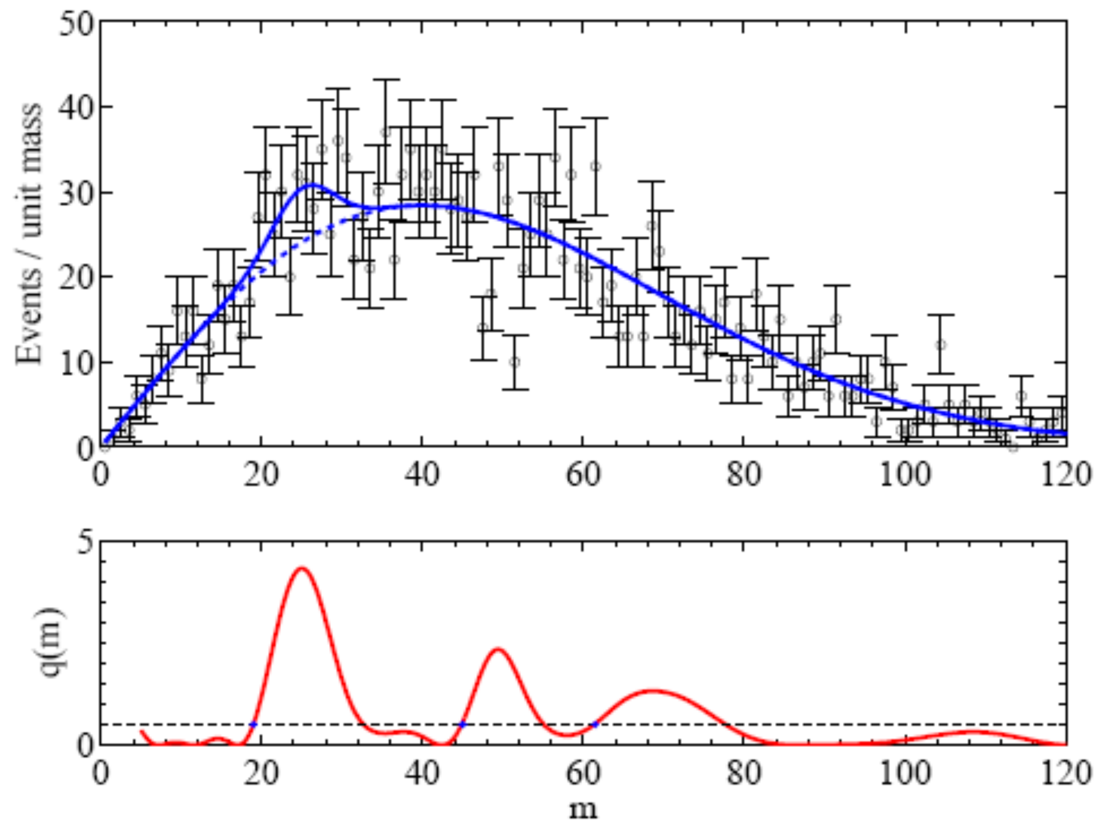
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c)\rangle$, the mean number "upcrossings" of $t_{\text{fix}} = -2\ln \lambda$ in the fit range based on a threshold $c = t_{\text{fix}} = Z_{\text{fix}}^2$.

$\langle N(c)\rangle$ can be estimated from MC (or the real data) using a much lower threshold c_0:

$$\langle N(c)\rangle \approx \langle N(c_0)\rangle e^{-(c-c_0)/2}$$

In this way $\langle N(c)\rangle$ can be estimated without need of large MC samples, even if the the threshold c is quite high.

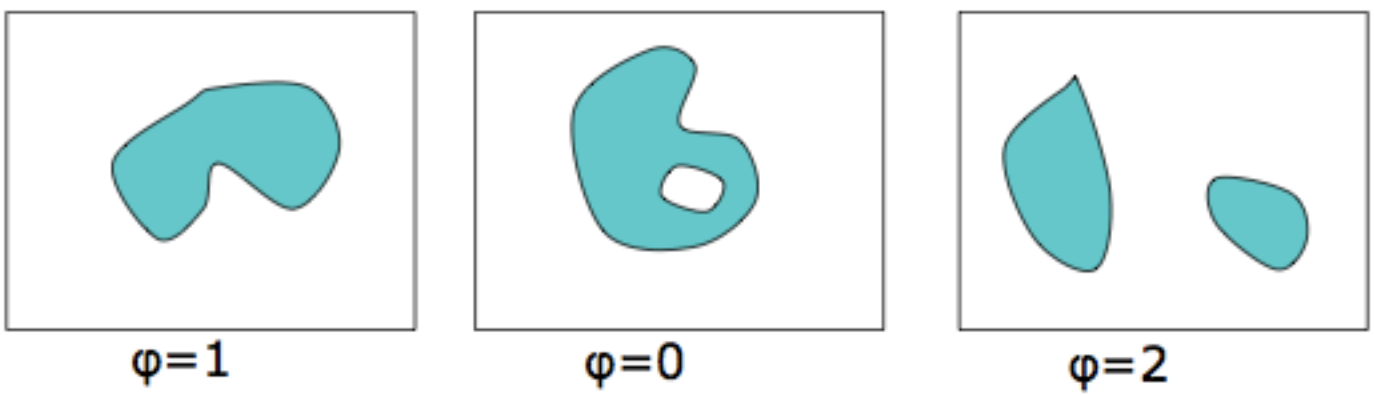


Multidimensional look-elsewhere effect

Generalization to multiple dimensions: number of upcrossings replaced by expectation of Euler characteristic:

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Number of disconnected components minus number of 'holes'



Applications: astrophysics (coordinates on sky), search for resonance of unknown mass and width, ...

Summary on Look-Elsewhere Effect

Remember the Look-Elsewhere Effect is when we test a single model (e.g., SM) with multiple observations, i..e, in mulitple places.

Note there is no look-elsewhere effect when considering exclusion limits. There we test specific signal models (typically once) and say whether each is excluded.

With exclusion there is, however, the analogous issue of testing many signal models (or parameter values) and thus excluding some even in the absence of signal (“spurious exclusion”)

Approximate correction for LEE should be sufficient, and one should also report the uncorrected significance.

“There's no sense in being precise when you don't even know what you're talking about.” — John von Neumann

Why 5 sigma?

Common practice in HEP has been to claim a discovery if the p-value of the no-signal hypothesis is below 2.9×10^{-7}, corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a 5σ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

The "cost" of announcing a false discovery is high.

Unsure about systematics.

Unsure about look-elsewhere effect.

The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).

Why 5 sigma (cont.)?

But the primary role of the *p*-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an "effect", and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to 3σ than 5σ.

Conclusions

Community relies largely on its own software tools:

ROOT, RooFit, RooStats, BAT, TMVA,...

Increasing use of multivariate methods: NN, BDT,...

Searches for new phenomena in particle physics have primarily been based on frequentist *p*-values.

Limited experience with e.g. Bayes factors but some movement in this direction (e.g. double beta decay; talk by Majorovits).

Situation with exclusion limits less stable:

one-sided frequentist limits,
unified (Feldman-Cousins) limits;
Bayesian limits (usually flat prior for signal rate);
CLs (limit based on penalized *p*-value).

Other issues: deconvolution (unfolding), measures of sensitivity,...

Extra slides

Test statistic for upper limits

cf. Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554.

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ:

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

Large sample approximation: $$\boxed{p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)}$$

95% CL upper limit on μ is highest value for which p-value is not less than 0.05.

Frequentist upper limit on Poisson parameter

Consider observing $n \sim \text{Poisson}(s + b)$

s, b = expected numbers of events from signal/background

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $\text{P}(n \leq n_{\text{obs}}; s, b)$

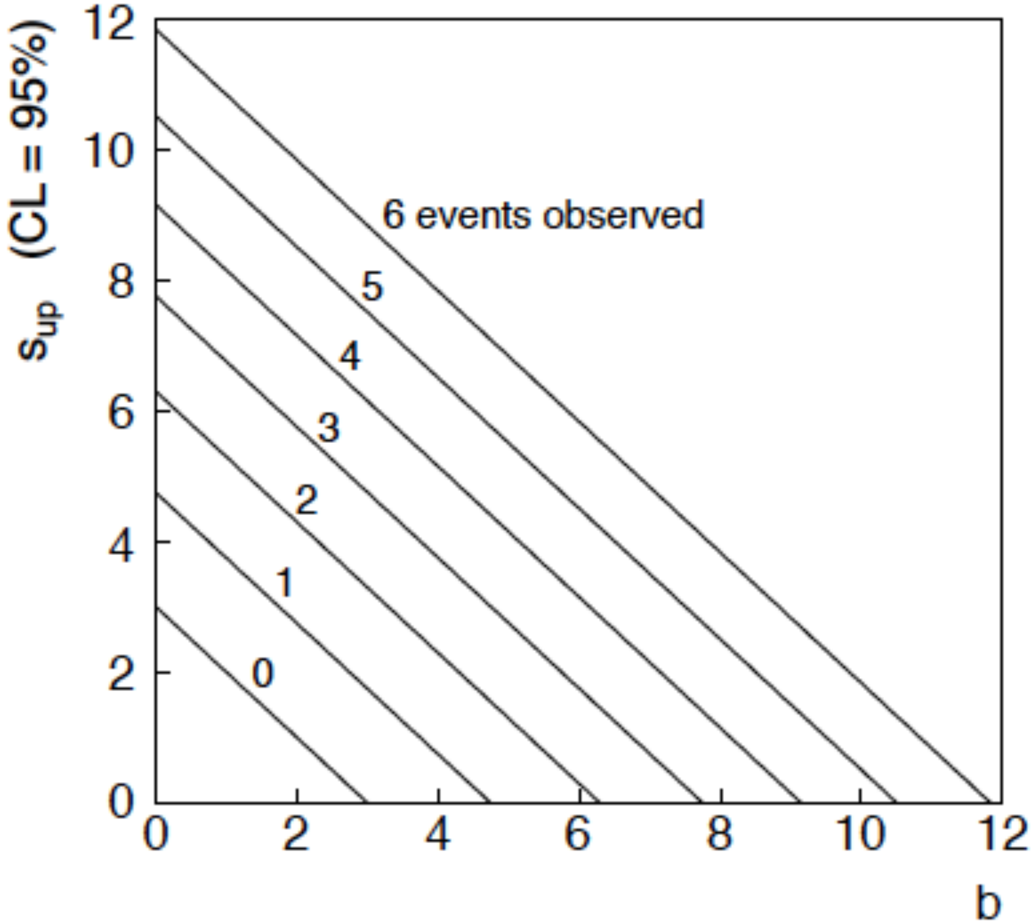
Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b \quad = 6.0$$

$n \sim$ Poisson($s+b$): frequentist upper limit on s

For low fluctuation of n formula can give negative result for s_{up}; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for $s_{\rm up}$

$$s_{\rm up} = -0.197 \quad ({\rm CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s.

Expected limit for $s = 0$

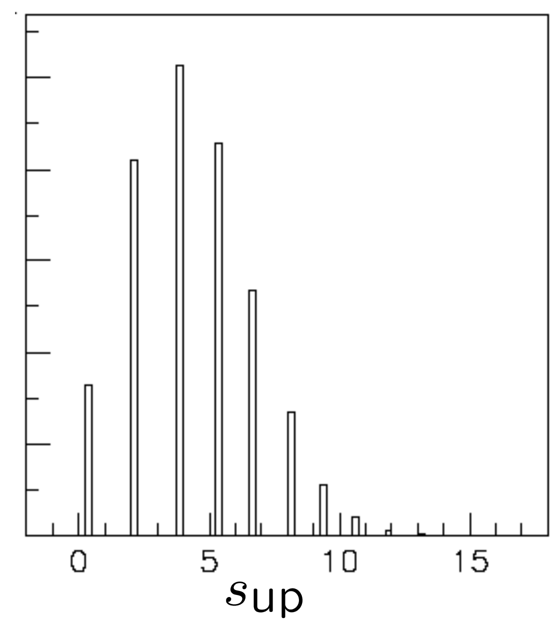
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta\,|\,x)$ to give interval with any desired probability content.

For e.g interval $[\theta_{\mathrm{lo}}, \theta_{\mathrm{up}}]$ with probability content $1 - \alpha$ one has

$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|x)\,d\theta$$

E.g., $\theta_{\mathrm{lo}} = -\infty$ for upper limit, $\theta_{\mathrm{up}} = +\infty$ for lower limit.

Bayesian prior for Poisson signal mean s

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian upper limit with flat prior for *s*

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto \frac{(s+b)^n}{n!} e^{-(s+b)} \qquad (s \geq 0)$$

Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete gamma function

Upper limit s_{up} determined by requiring

$$1 - \alpha = \int_0^{s_{\rm up}} p(s|n)\, ds$$

Bayesian interval with flat prior for *s*

Solve to find limit s_{up}:

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} \left[p, 2(n+1) \right] - b$$

where

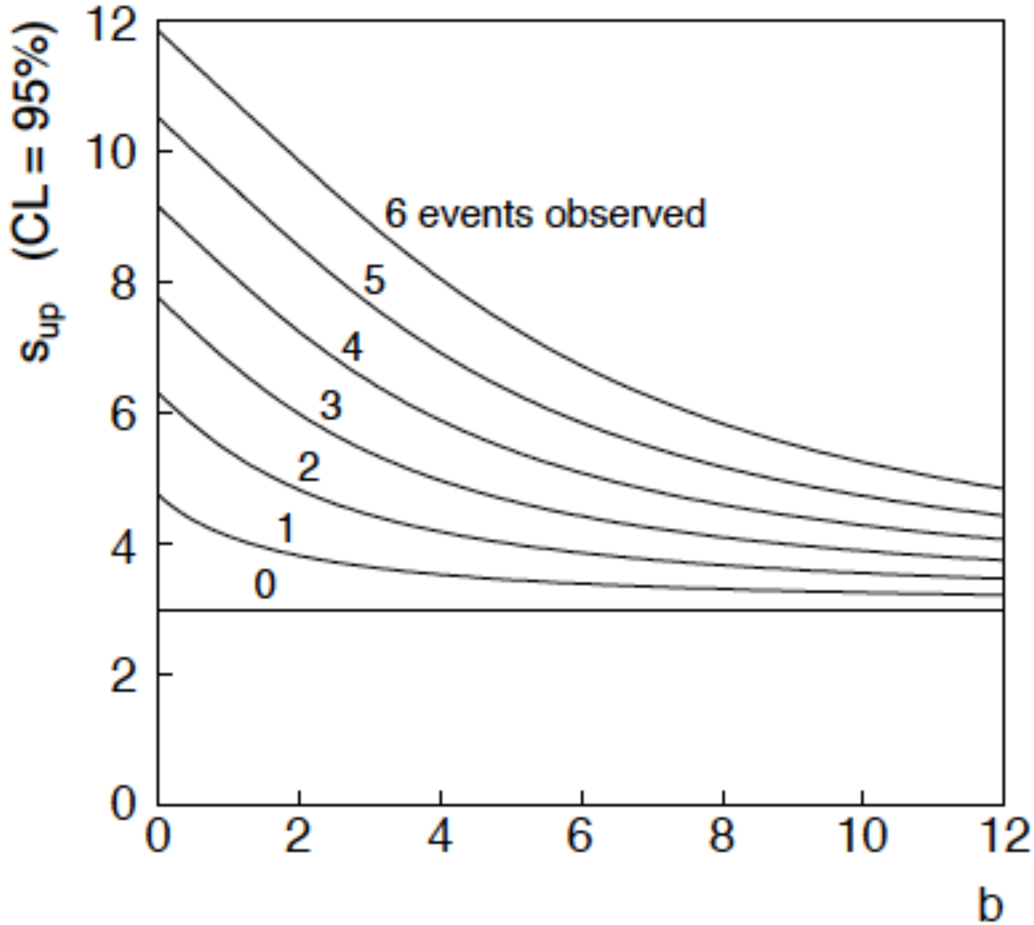
$$p = 1 - \alpha \left(1 - F_{\chi^2} \left[2b, 2(n+1) \right] \right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case (‘coincidence’).

Bayesian interval with flat prior for s

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with certain coverage properties.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

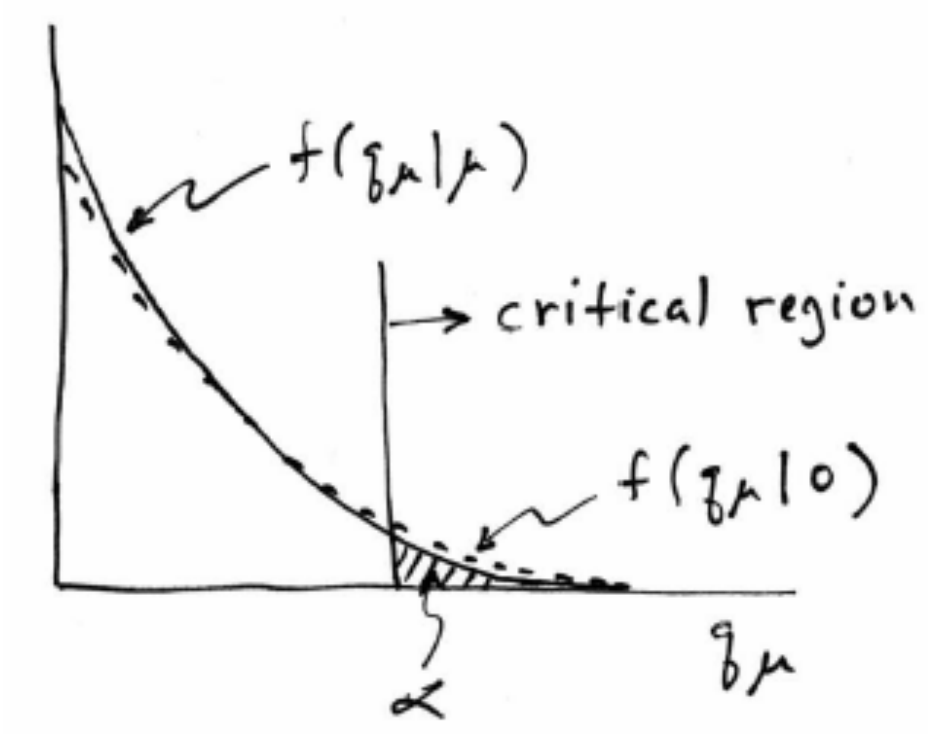
L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Low sensitivity to μ

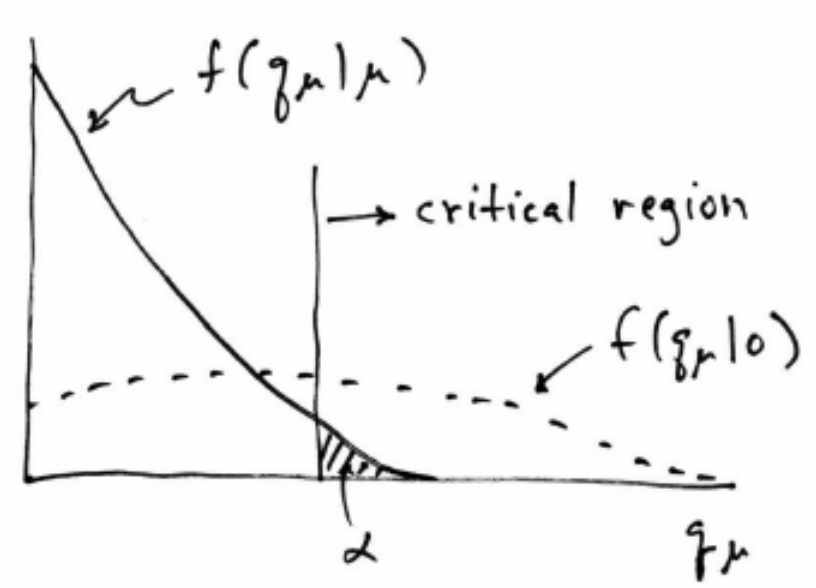
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

This means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ will be almost the same:



Having sufficient sensitivity

In contrast, having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are more separated:

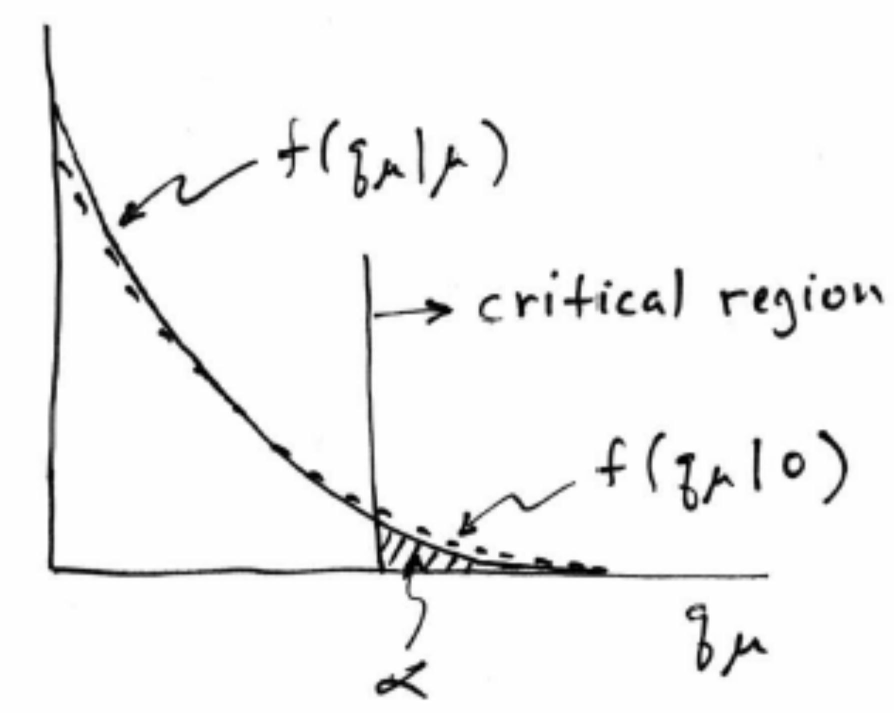


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α. Use this power as a measure of the sensitivity.

Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject μ if μ is true is α (e.g., 5%).

And the probability to reject μ if $\mu = 0$ (the power) is only slightly greater than α.



This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_H = 1000$ TeV).

"Spurious exclusion"

Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

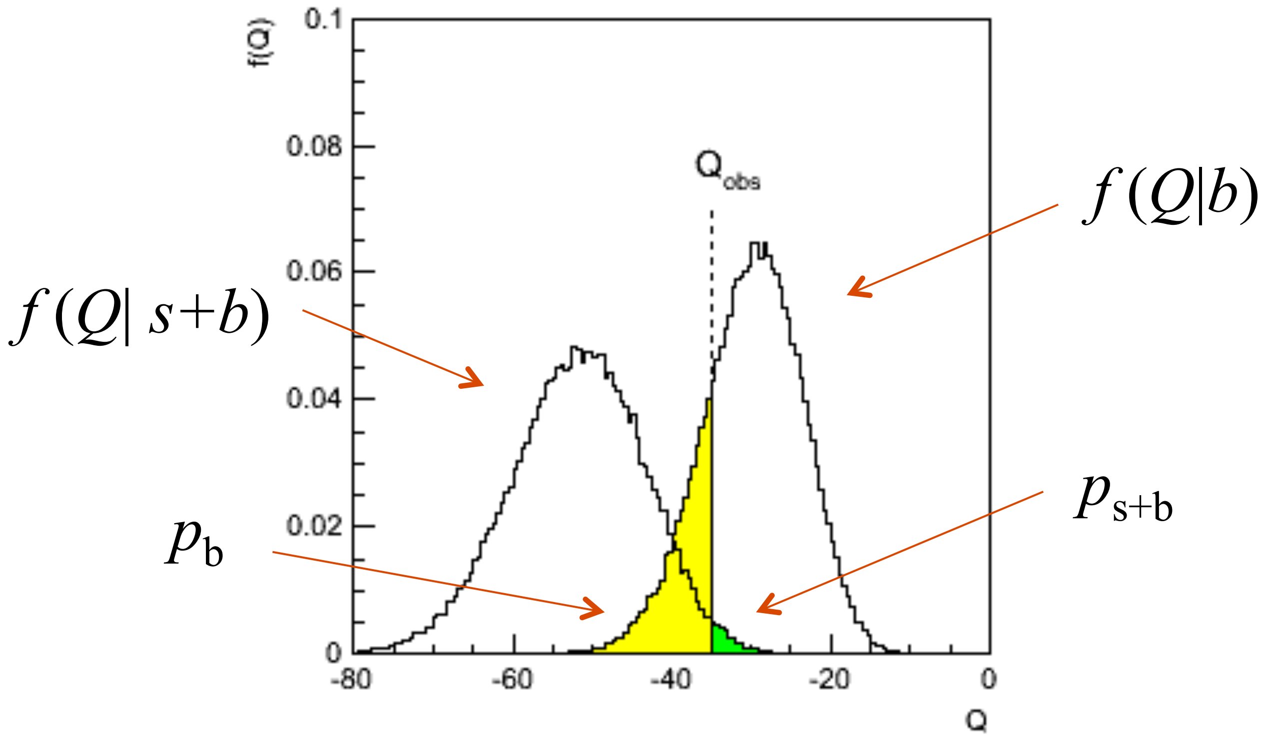
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the "CL$_s$" procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.

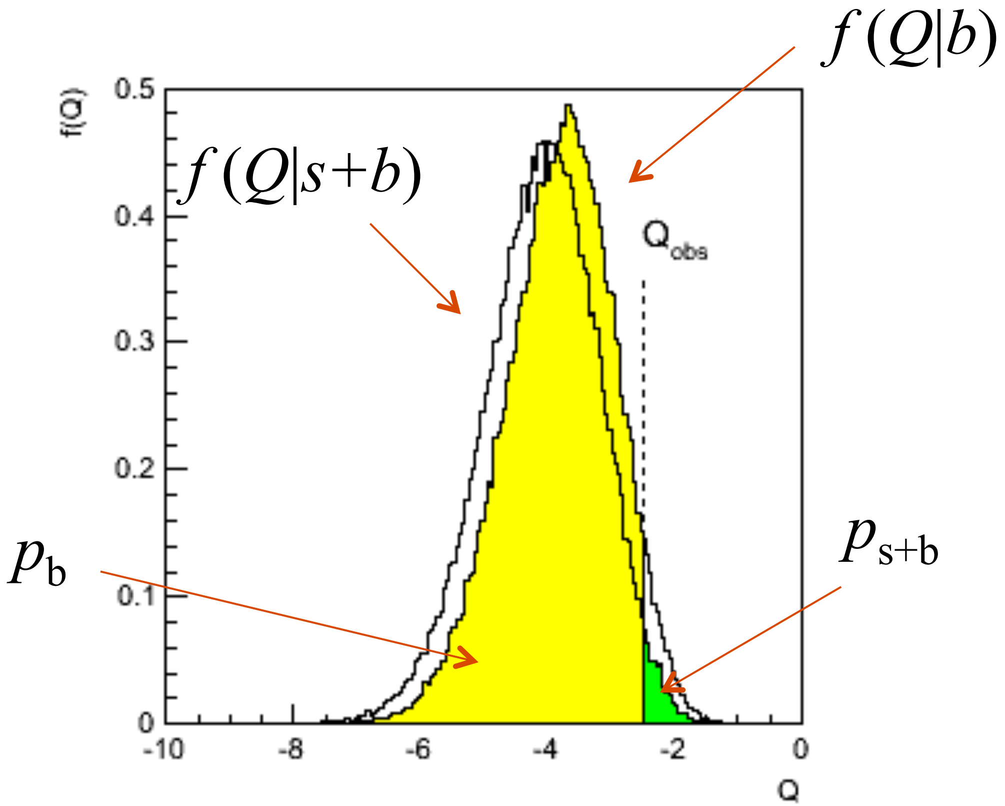
The CL$_s$ procedure

In the usual formulation of CL$_s$, one tests both the $\mu = 0$ (b) and $\mu > 0$ ($\mu s+b$) hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$:



The CL$_s$ procedure (2)

As before, "low sensitivity" means the distributions of Q under b and $s+b$ are very close:



The CL$_s$ procedure (3)

The CL$_s$ solution (A. Read et al.) is to base the test not on the usual *p*-value (CL$_{s+b}$), but rather to divide this by CL$_b$ (~ one minus the *p*-value of the *b*-only hypothesis), i.e.,

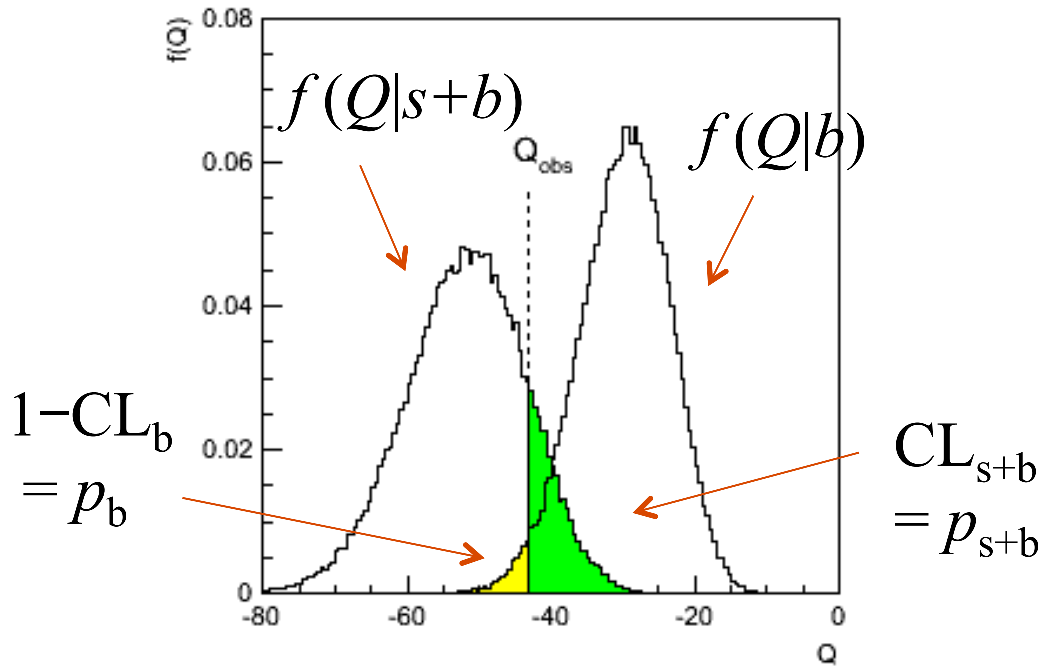
Define:

$$\mathrm{CL_s} = \frac{\mathrm{CL_{s+b}}}{\mathrm{CL_b}}$$

$$= \frac{p_{s+b}}{1 - p_b}$$

Reject s+b hypothesis if:

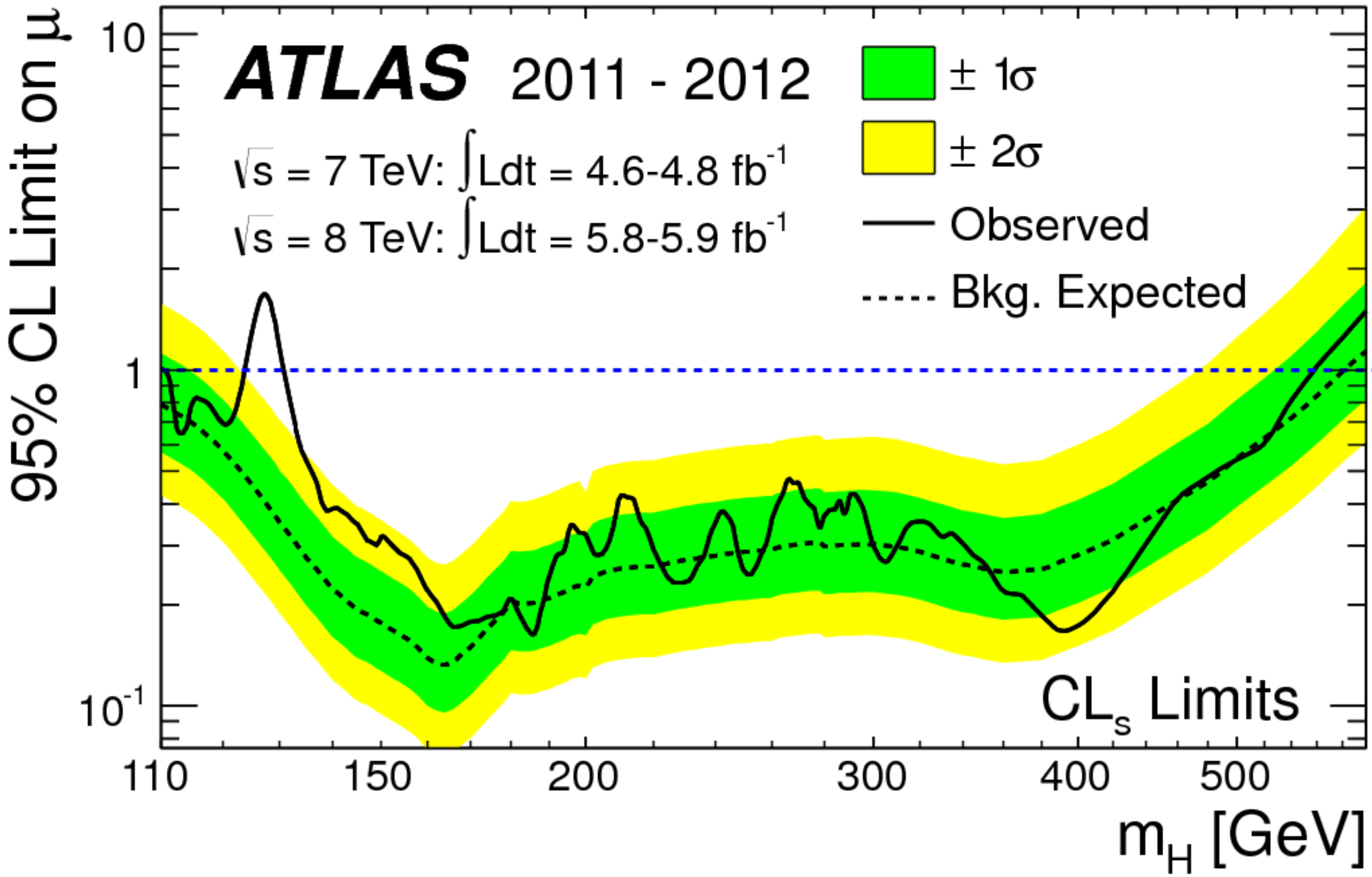
$$\mathrm{CL_s} \leq \alpha$$



Increases "effective" *p*-value when the two distributions become close (prevents exclusion if sensitivity is low).

Example of CLs upper limit

ATLAS Collaboration, Phys. Lett. B 716 (2012) 1-29



Choice of test for limits (2)

In some cases $\mu = 0$ is no longer a relevant alternative and we want to try to exclude μ on the grounds that some other measure of incompatibility between it and the data exceeds some threshold.

If the measure of incompatibility is taken to be the likelihood ratio with respect to a two-sided alternative, then the critical region can contain both high and low data values.

→ unified intervals, G. Feldman, R. Cousins, Phys. Rev. D 57, 3873–3889 (1998)

The Big Debate is whether to use one-sided or unified intervals in cases where small (or zero) values of the parameter are relevant alternatives. Professional statisticians have voiced support on both sides of the debate.

Unified (Feldman-Cousins) intervals

We can use directly

$$t_\mu = -2 \ln \lambda(\mu) \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

as a test statistic for a hypothesized μ.

Large discrepancy between data and hypothesis can correspond either to the estimate for μ being observed high or low relative to μ.

This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873.

Lower edge of interval can be at $\mu = 0$, depending on data.

Distribution of t_μ

Using Wald approximation, $f(t_\mu|\mu')$ is noncentral chi-square for one degree of freedom:

$$f(t_\mu|\mu') = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2} \left(\sqrt{t_\mu} + \frac{\mu - \mu'}{\sigma} \right)^2 \right) + \exp\left(-\frac{1}{2} \left(\sqrt{t_\mu} - \frac{\mu - \mu'}{\sigma} \right)^2 \right) \right]$$

Special case of $\mu = \mu'$ is chi-square for one d.o.f. (Wilks).

The p-value for an observed value of t_μ is

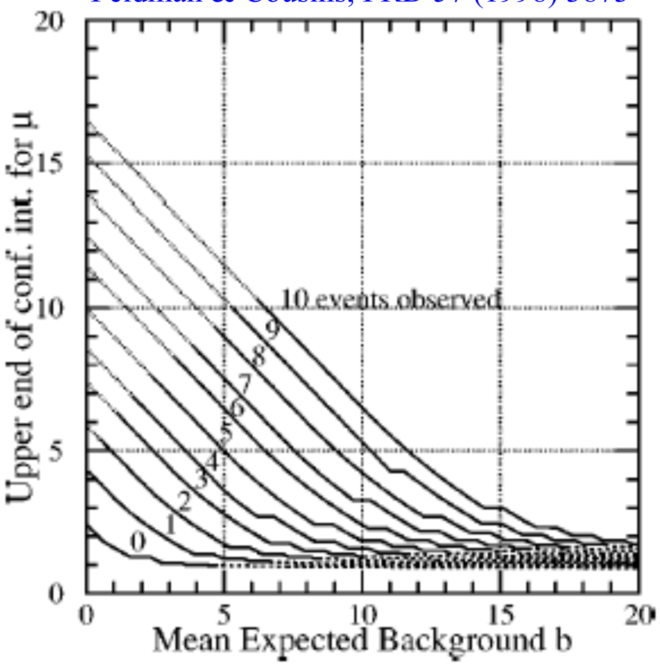
$$p_\mu = 1 - F(t_\mu|\mu) = 2\left(1 - \Phi\left(\sqrt{t_\mu}\right)\right)$$

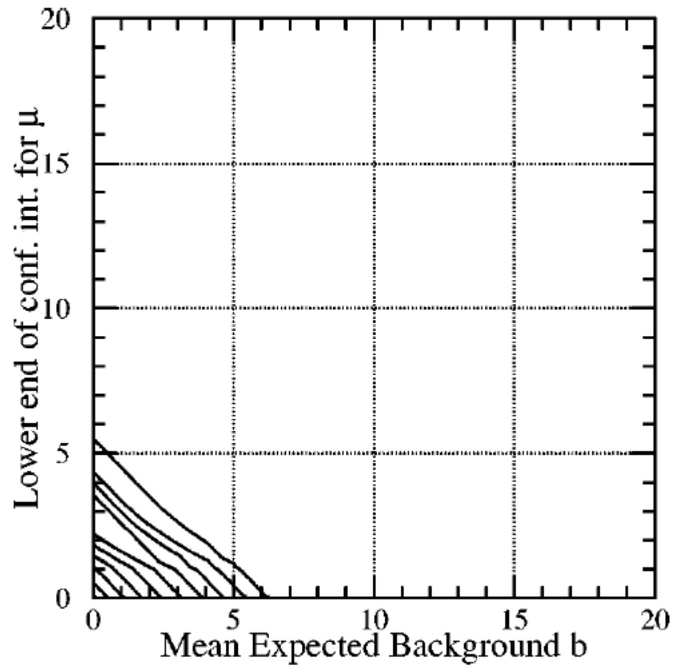
and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \Phi^{-1}\left(2\Phi\left(\sqrt{t_\mu}\right) - 1\right)$$

Upper/lower edges of F-C interval for μ versus b for $n \sim \text{Poisson}(\mu+b)$

Feldman & Cousins, PRD 57 (1998) 3873





Lower edge may be at zero, depending on data.

For $n = 0$, upper edge has (weak) dependence on b.

Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

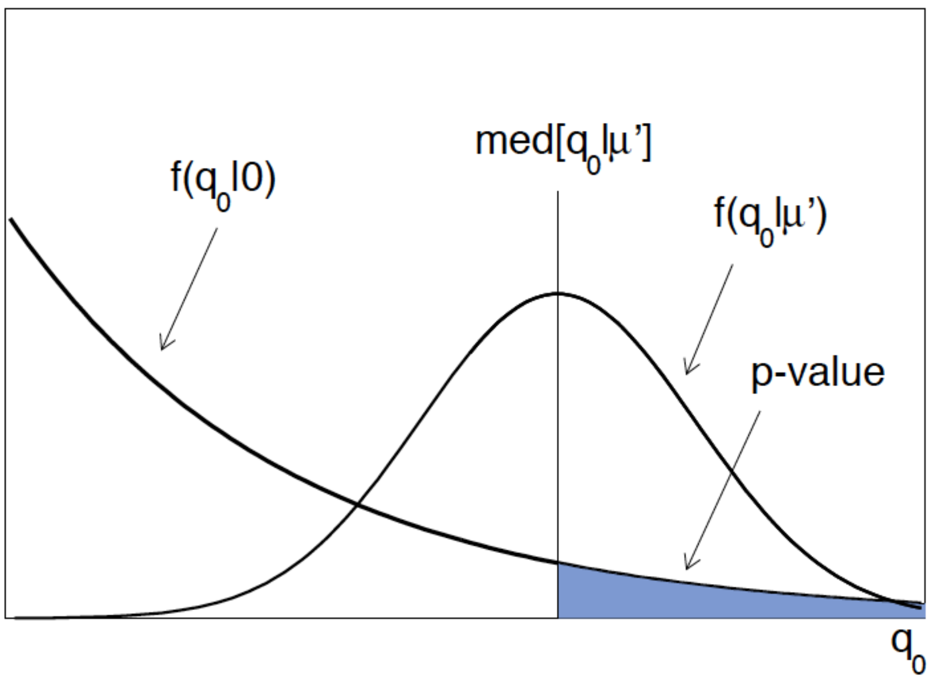
The F-C limits are based on a likelihood ratio for a test of μ with respect to the alternative consisting of all other allowed values of μ (not just, say, lower values).

The interval's upper edge is higher than the limit from the one-sided test, and lower values of μ may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of μ is excluded, it is because there is a probability α for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.

Expected (or median) significance / sensitivity

When planning the experiment, we want to quantify how sensitive we are to a potential discovery, e.g., by given median significance assuming some nonzero strength parameter μ'.



So for p-value, need $f(q_0|0)$, for sensitivity, will need $f(q_0|\mu')$,

Back to Poisson counting experiment

$n \sim$ Poisson(s+b), where

s = expected number of events from signal,

b = expected number of background events.

To test for discovery of signal compute p-value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\rm obs}|b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\rm obs})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1-p)$
where Φ is the standard Gaussian cumulative distribution, e.g.,
$Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s.

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \rightarrow x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{obs} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\rm obs} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\rm obs} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\mathrm{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter s is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

For now no nuisance params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{s} \geq 0\,, \\ 0 & \hat{s} < 0\,. \end{cases} \qquad \lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2\ln\frac{L(0)}{L(\hat{s})} = 2\left(n\ln\frac{n}{b} + b - n\right) \quad \text{for } n > b, \text{ 0 otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \qquad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

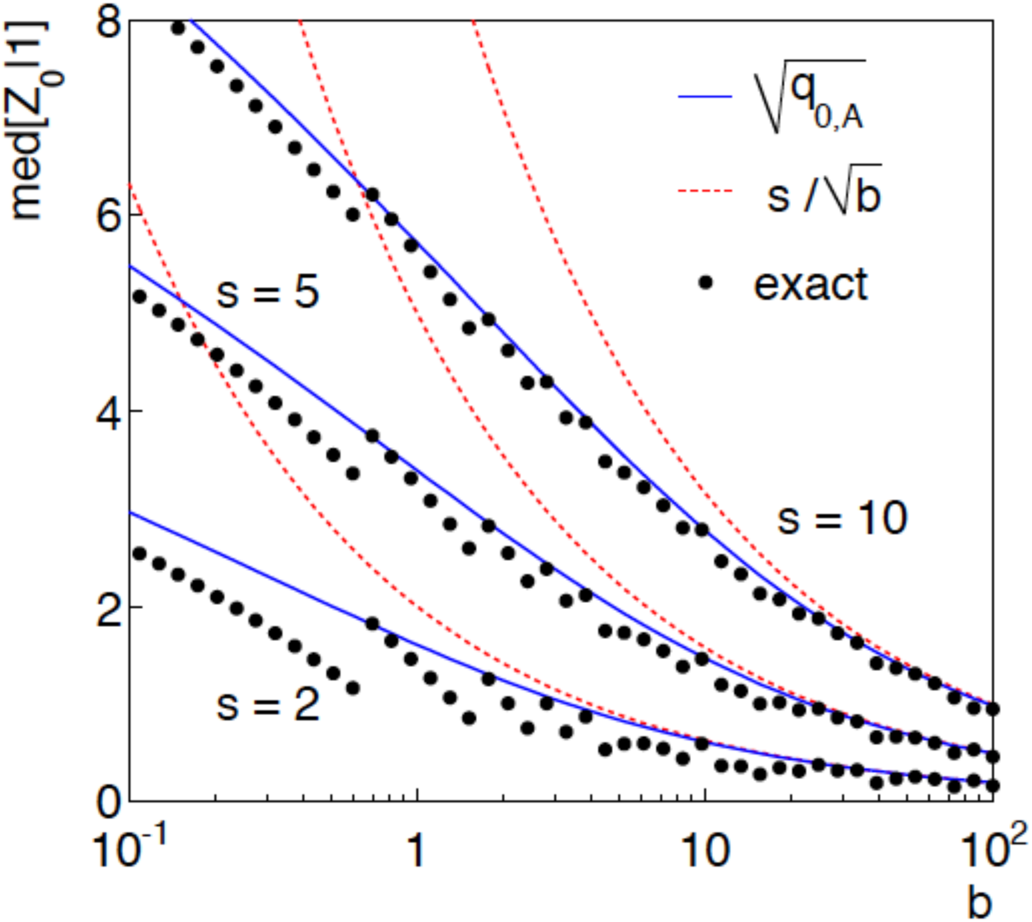
To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

$$Z_{\mathrm{A}} = \sqrt{2 \left((s+b) \ln \left(1 + \frac{s}{b} \right) - s \right)}$$

This reduces to $s/\sqrt{b}$ for s << b.

$n \sim$ Poisson(s+b), median significance, assuming s, of the hypothesis s = 0

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,A}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.

Extending $s/\sqrt{b}$ to case where b uncertain

The intuitive explanation of $s/\sqrt{b}$ is that it compares the signal, s, to the standard deviation of n assuming no signal, $\sqrt{b}$.

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b.

A reasonable guess is to replace $\sqrt{b}$ by the quadratic sum of $\sqrt{b}$ and σ_b, i.e.,

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Adding a control measurement for b

(The "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...)

Measure two Poisson distributed values:

$n \sim \text{Poisson}(s+b)$ (primary or "search" measurement)

$m \sim \text{Poisson}(\tau b)$ (control measurement, τ known)

The likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct profile likelihood ratio (b is nuisance parmeter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\begin{aligned}
\hat{s} &= n - m/\tau \,, \\
\hat{b} &= m/\tau \,, \\
\hat{\hat{b}}(s) &= \frac{n + m - (1+\tau)s + \sqrt{(n + m - (1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} \,.
\end{aligned}$$

and in particular to test for discovery ($s = 0$),

$$\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}$$

Asymptotic significance

Use profile likelihood ratio for q_0, and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0}$$

$$= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n, m by their expectation values assuming background-plus-signal model:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

$$Z_{\mathrm{A}} = \left[-2 \left((s+b) \ln \left[\frac{s + (1+\tau)b}{(1+\tau)(s+b)} \right] + \tau b \ln \left[1 + \frac{s}{(1+\tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \dfrac{b}{\tau}$, to eliminate τ:

$$\boxed{Z_{\mathrm{A}} = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}}$$

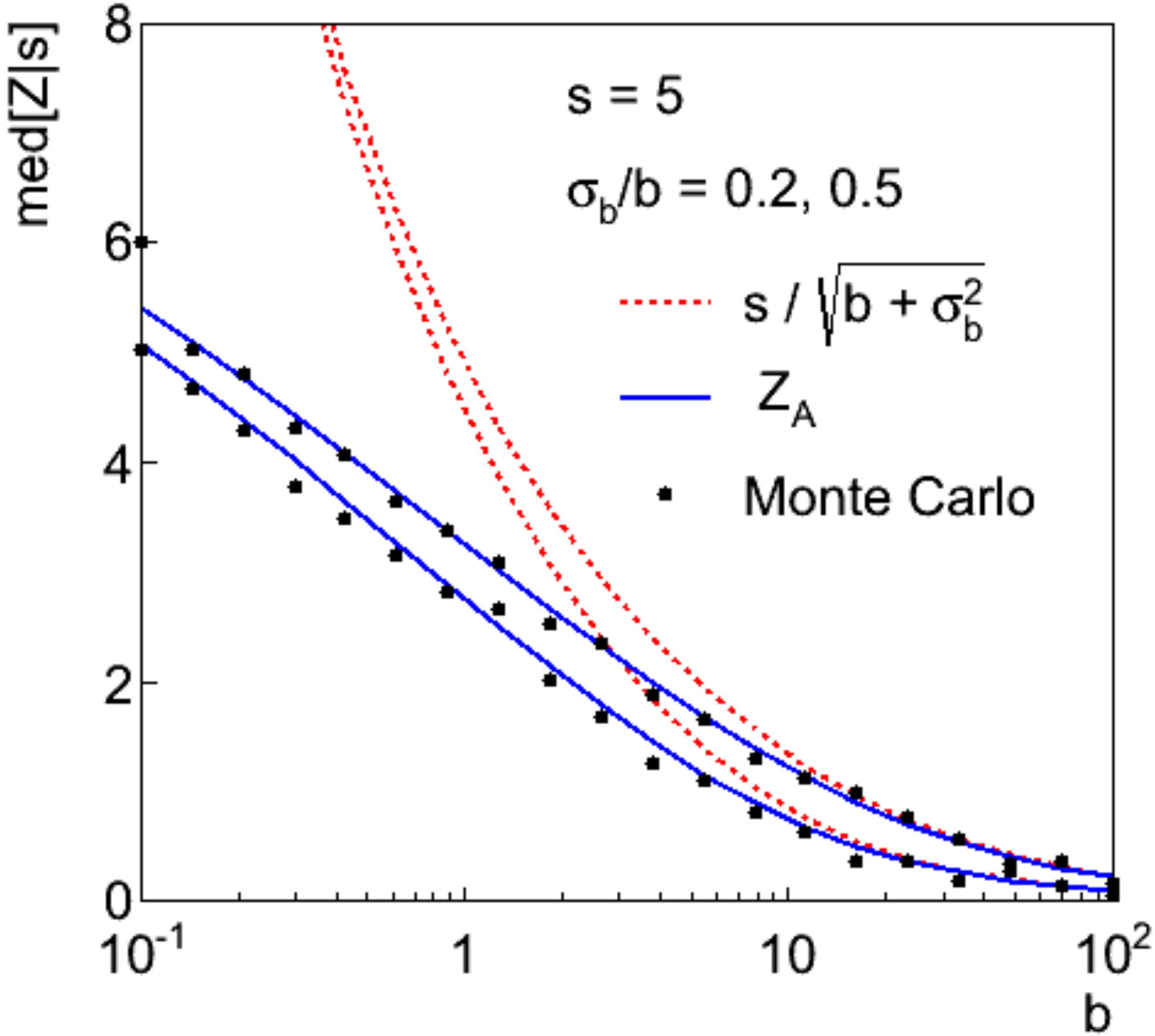
Limiting cases

Expanding the Asimov formula in powers of s/b and σ_b^2/b (= $1/\tau$) gives

$$Z_{\mathrm{A}} = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So this "intuitive" formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$



Using sensitivity to optimize a cut

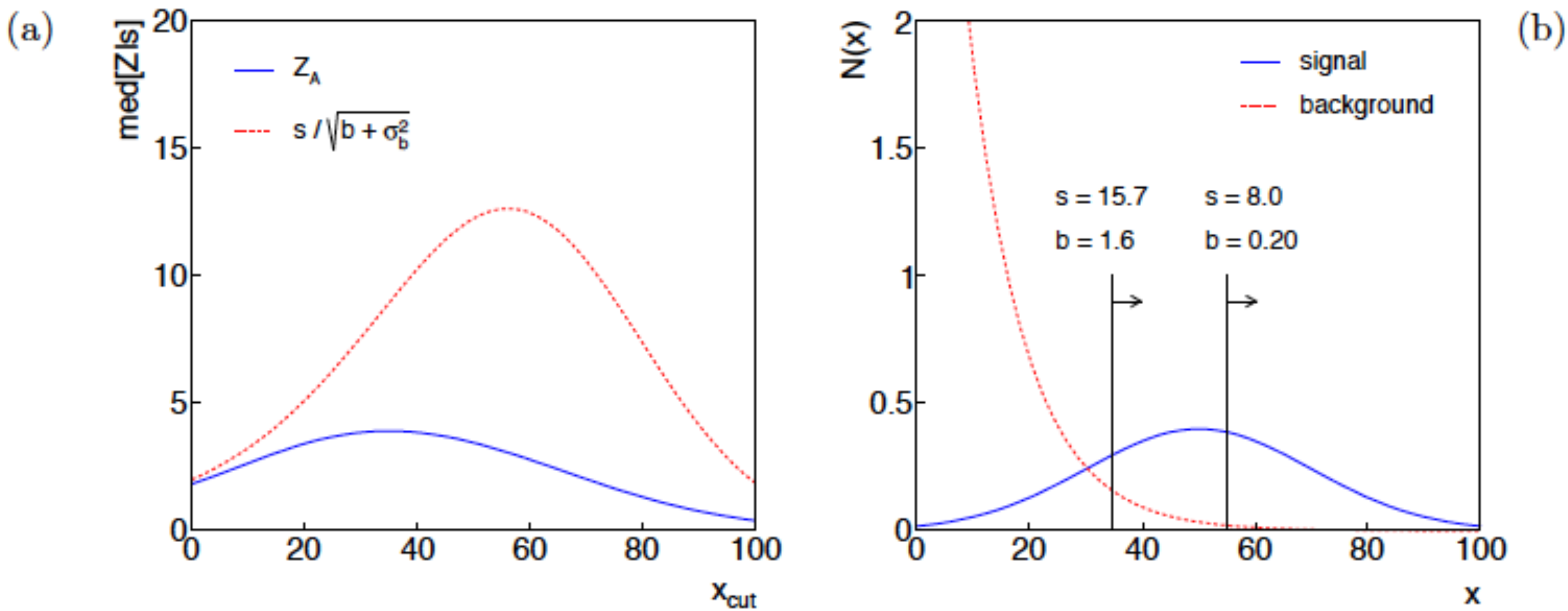
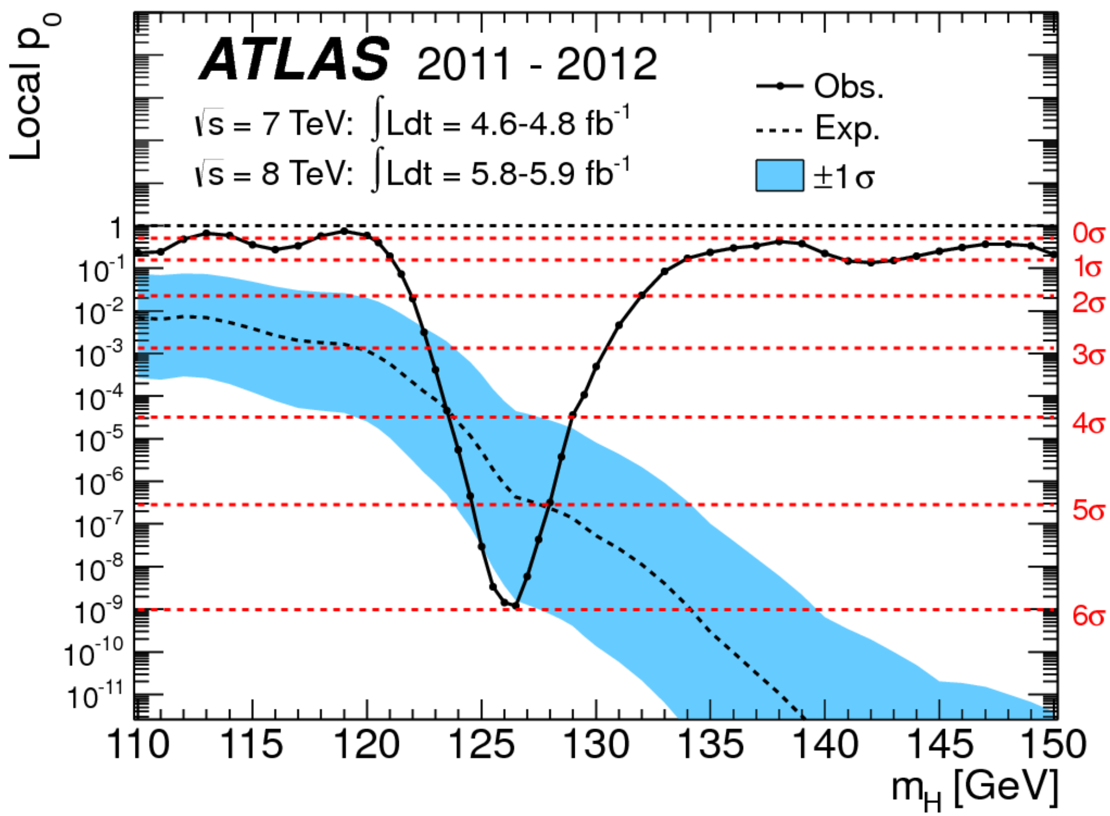


Figure 1: (a) The expected significance as a function of the cut value x_{cut}; (b) the distributions of signal and background with the optimal cut value indicated.

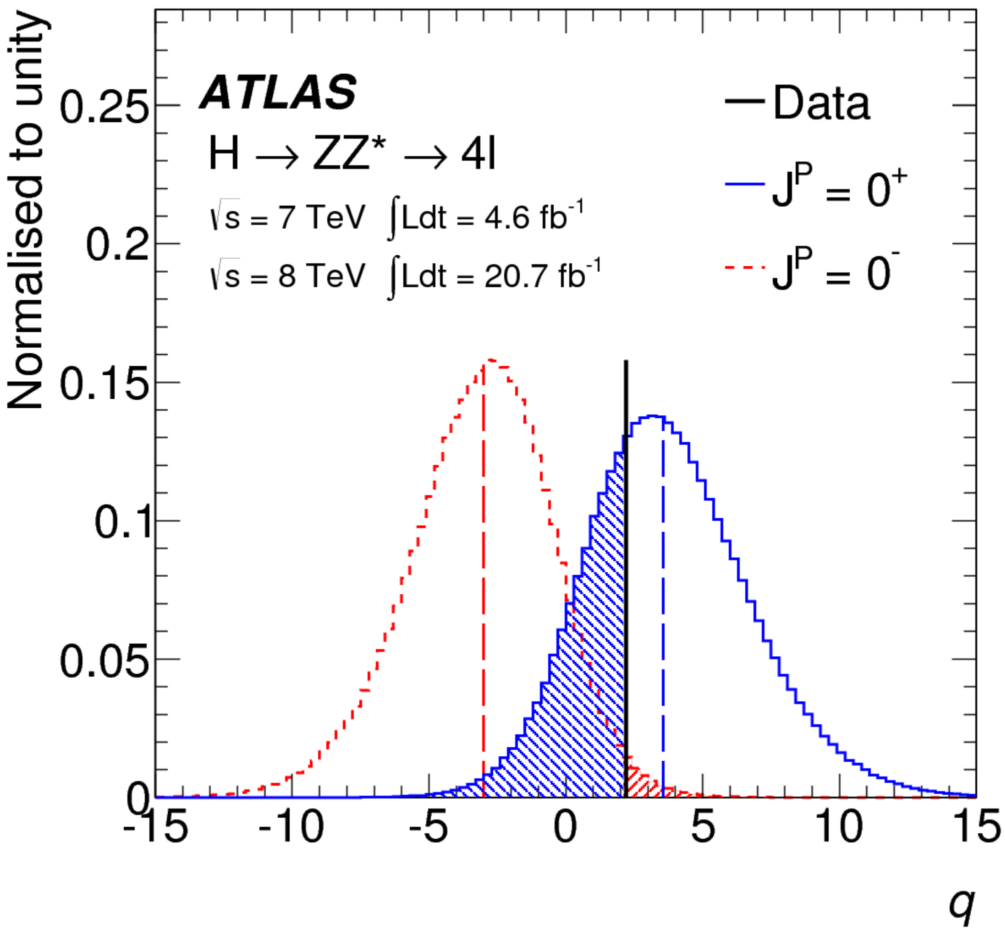
p-value for Higgs discovery

ATLAS Collaboration, Phys. Lett. B 716 (2012) 1-29



Comparison of two point hypotheses

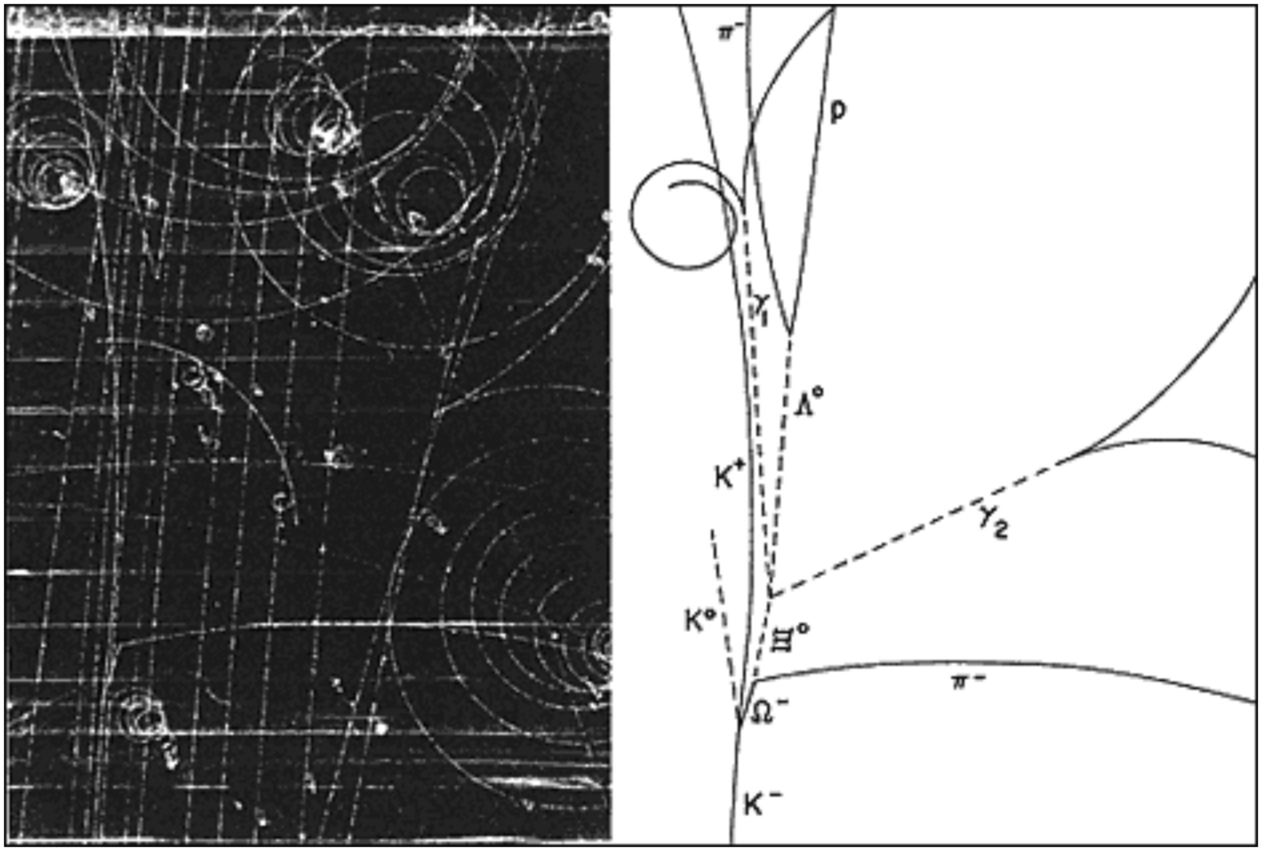
Evidence for the spin-0 nature of the Higgs boson using ATLAS data, ATLAS, Phys. Lett. B 726 (2013), pp. 120-144



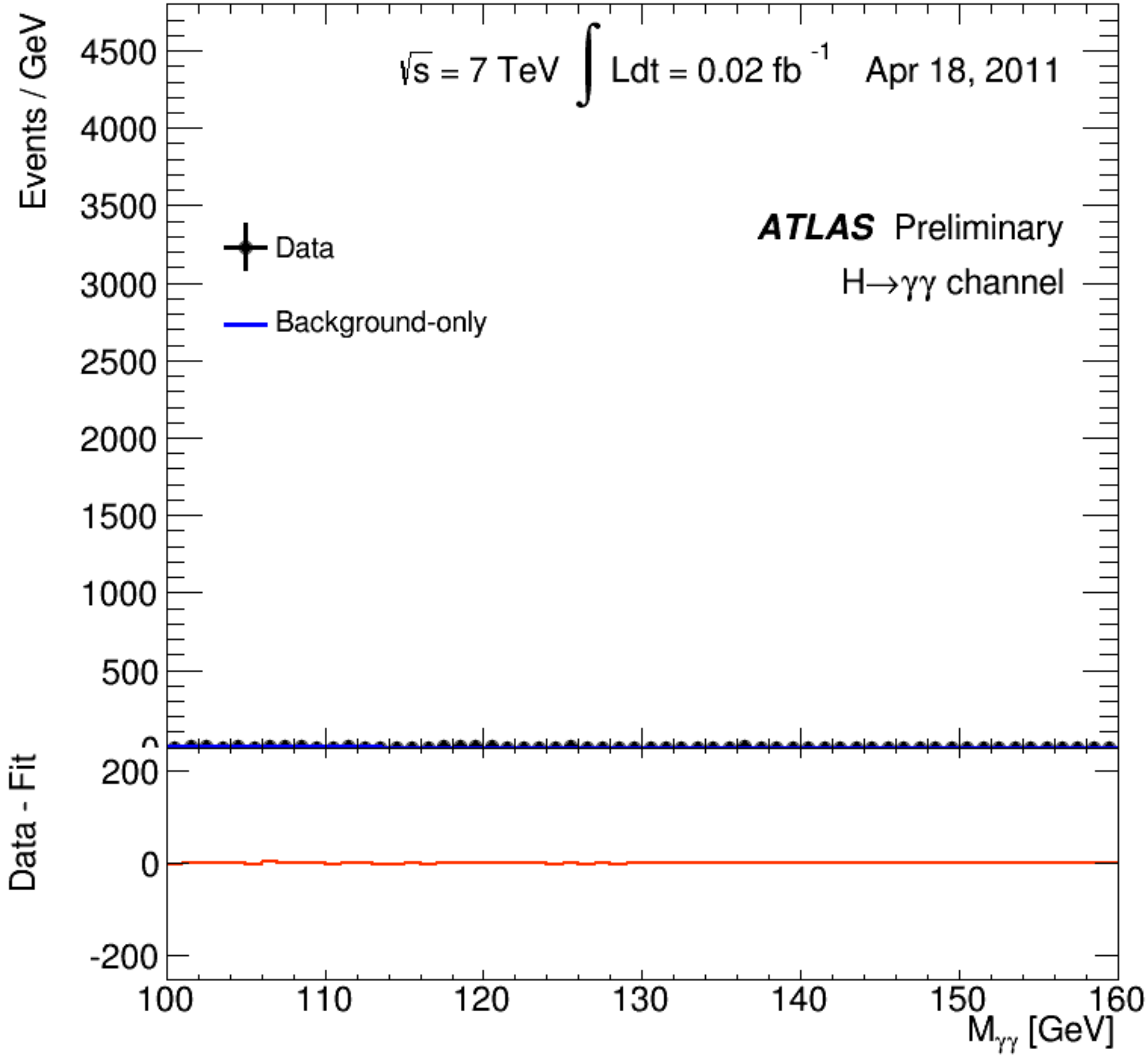
Result is summarized as *two p*-values, one for each model.

Sometimes discovery comes in a single event

Discovery of the Omega minus baryon, Samios et al., 1961

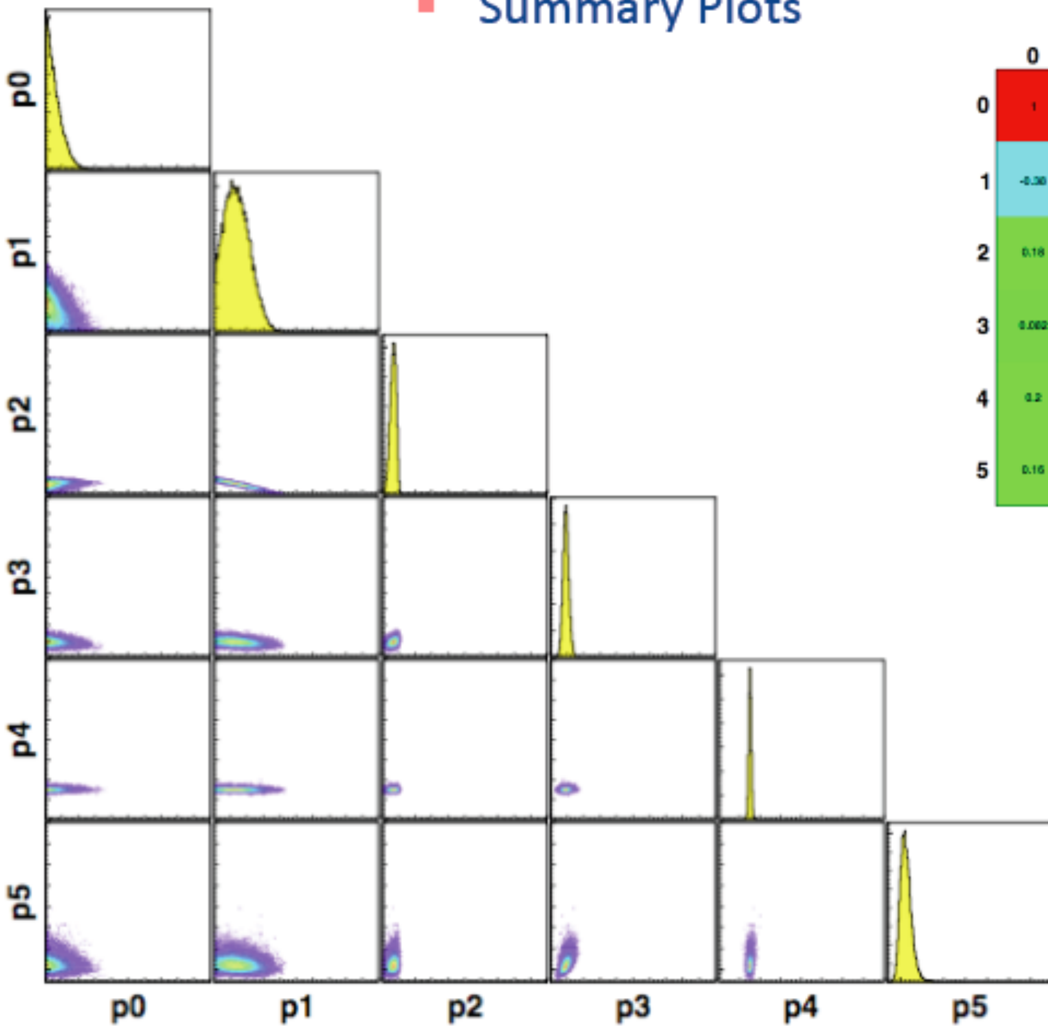


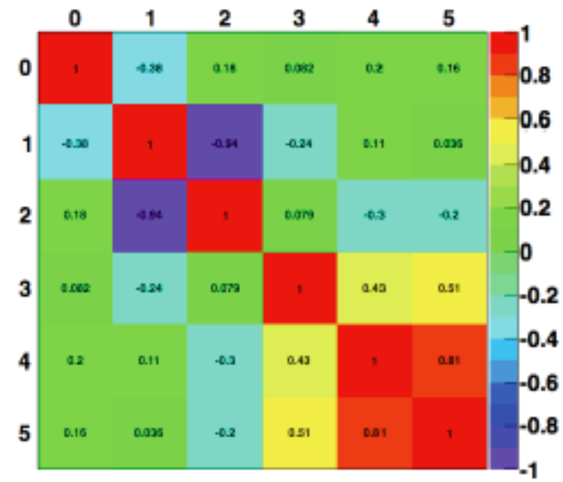
Usually discoveries appear gradually

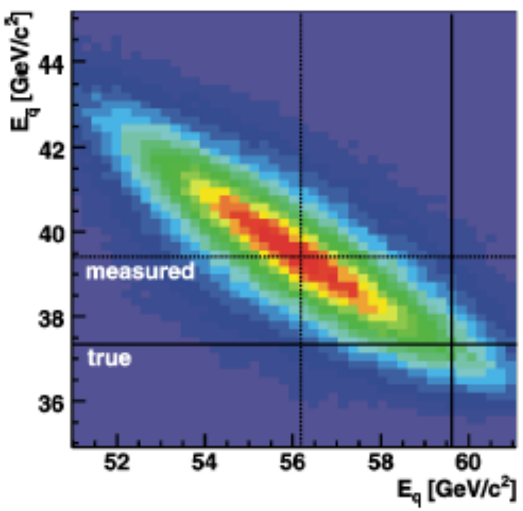


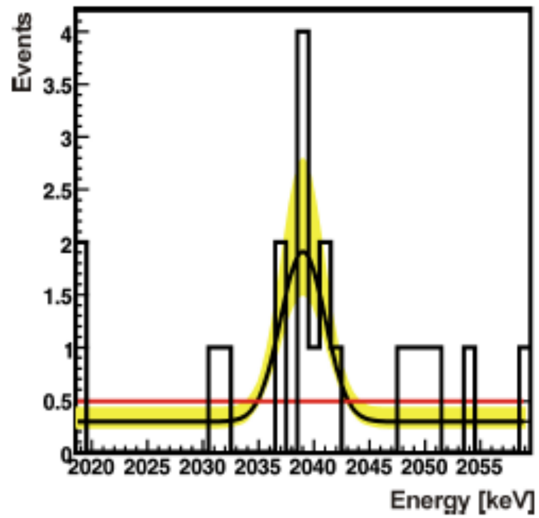
BAT MCMC Summary Plots

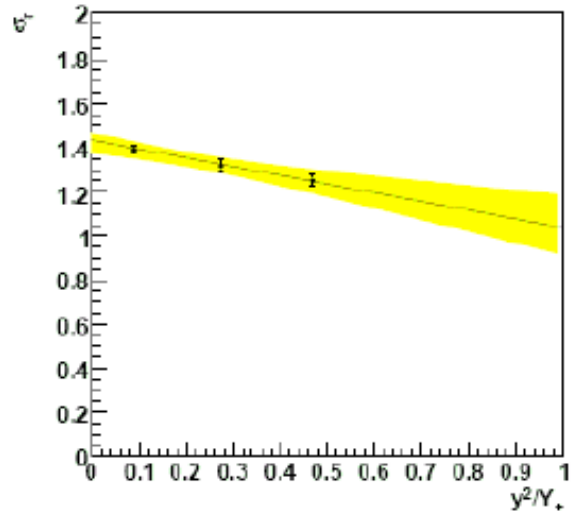
- Summary Plots











- Quentin Buat, *Search for extra dimensions in the diphoton final state with ATLAS* [arXiv:1201.4748]
- ATLAS collaboration, *Search for excited leptons in proton-proton collisions at sqrt(s) = 7 TeV with the ATLAS detector* [arXiv:1201.3293]
- I. Abt *et al.*, *Measurement of the temperature dependence of pulse lengths in an n-type germanium detector*, Eur. Phys. J. Appl. Phys.56:10104,2011 [arXiv:1112.5033]
- ATLAS collaboration, *Search for Extra Dimensions using diphoton events in 7 TeV proton-proton collisions with the ATLAS detector* [arXiv:1112.2194]
- ATLAS collaboration, *A measurement of the ratio of the W and Z cross sections with exactly one associated jet in pp collisions at sqrt(s) = 7 TeV with ATLAS*, Phys.Lett.B708:221-240,2012 [arXiv:1108.4908]
- ZEUS collaboration, *Search for single-top production in ep collisions at HERA*, Phys.Lett.B708:27-36,2012 [arXiv:1111.3901]
- CMS collaboration, *Search for a W' boson decaying to a muon and a neutrino in pp collisions at sqrt(s) = 7 TeV*, Phys.Lett.B701:160-179,2011 [arXiv:1103.0030]
- ZEUS collaboration, *Measurement of the Longitudinal Proton Structure Function at HERA*, Phys.Lett.B682:8-22,2009 [arXiv:0904.1092]

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form, also impossible with 'normal' acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$ e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha,\quad \vec{\theta}_1 = \vec{\theta}$, $\leftarrow$ move to proposed point

else $\quad \vec{\theta}_1 = \vec{\theta}_0$ $\leftarrow$ old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

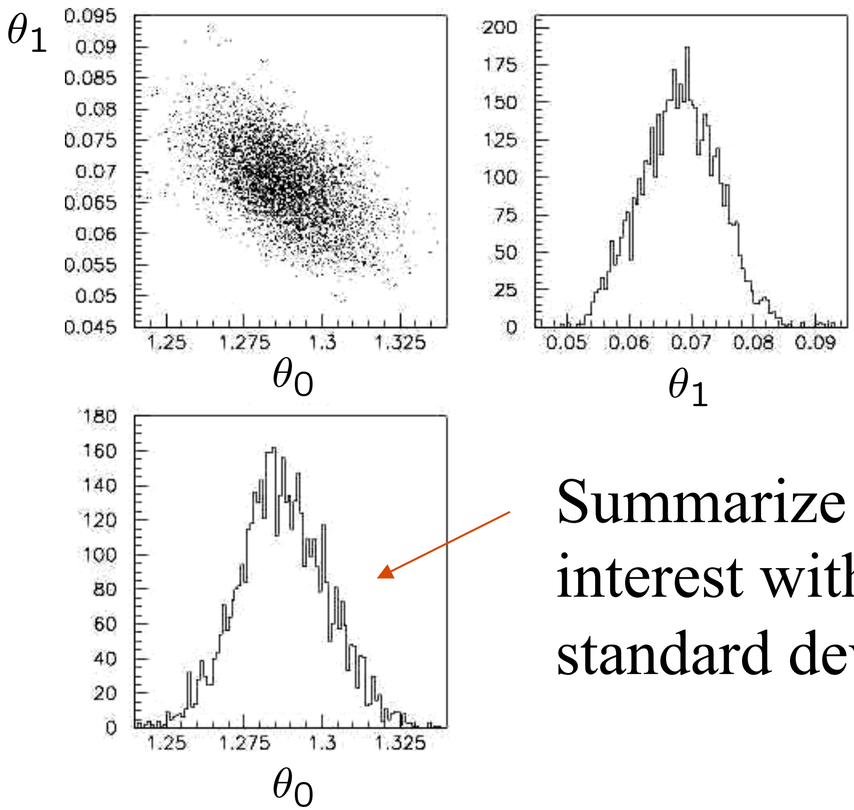
The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.
If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

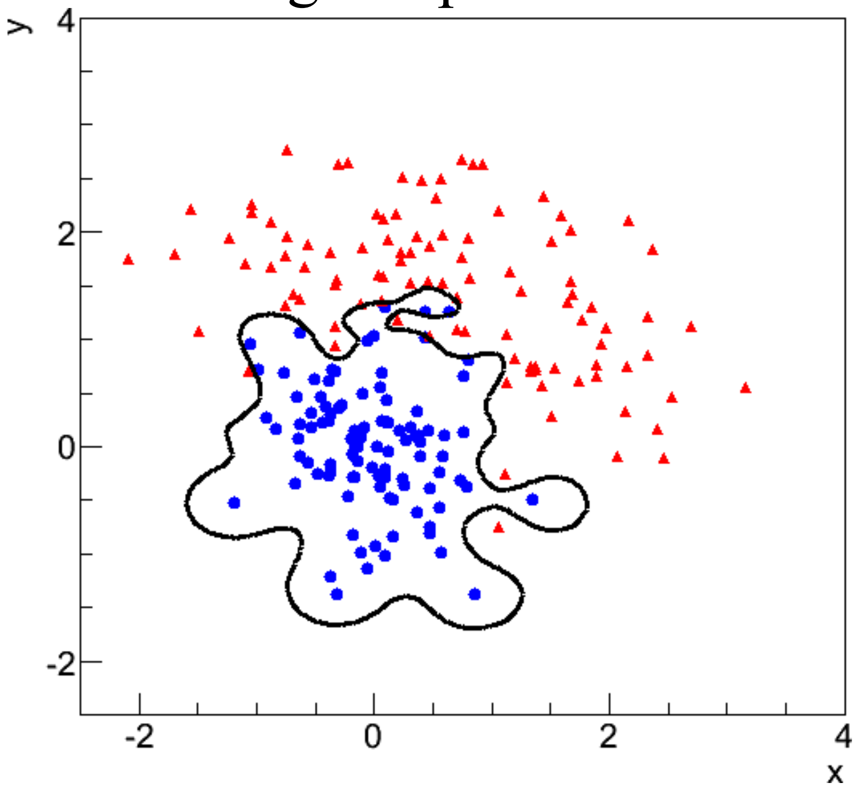
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Overtraining

If decision boundary is too flexible it will conform too closely to the training points → overtraining.

Monitor by applying classifier to independent validation sample.

training sample



independent validation sample