Statistical Methods for the LHC

Discovering New Physics

PSI Zuoz Summer School
1-7 August, 2010

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

Outline

Formalism of a search as a statistical test

Multivariate methods for statistical tests

Example: the Boosted Decision Tree

Using a test statistic for Discovery and setting limits

Significance

Sensitivity

Including systematics

Conclusions

Quick review of probablility

Frequentist (*A* = outcome of repeatable observation):

$$P(A) = \lim_{n \to \infty} \frac{\text{outcome is } A}{n}$$

Subjective (*A* = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Hypotheses

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim f(x|H)$.

x could represent e.g. observation of a single particle, a single event, or an entire “experiment”.

Possible values of x form the sample space S (or “data space”).

Simple (or “point”) hypothesis: $f(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Definition of a test

Consider e.g. a simple hypothesis H_0 and alternative H_1.

A test of H_0 is defined by specifying a critical region W of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W \mid H_0) \leq \alpha$$

If x is observed in the critical region, reject H_0.

α is called the size or significance level of the test.

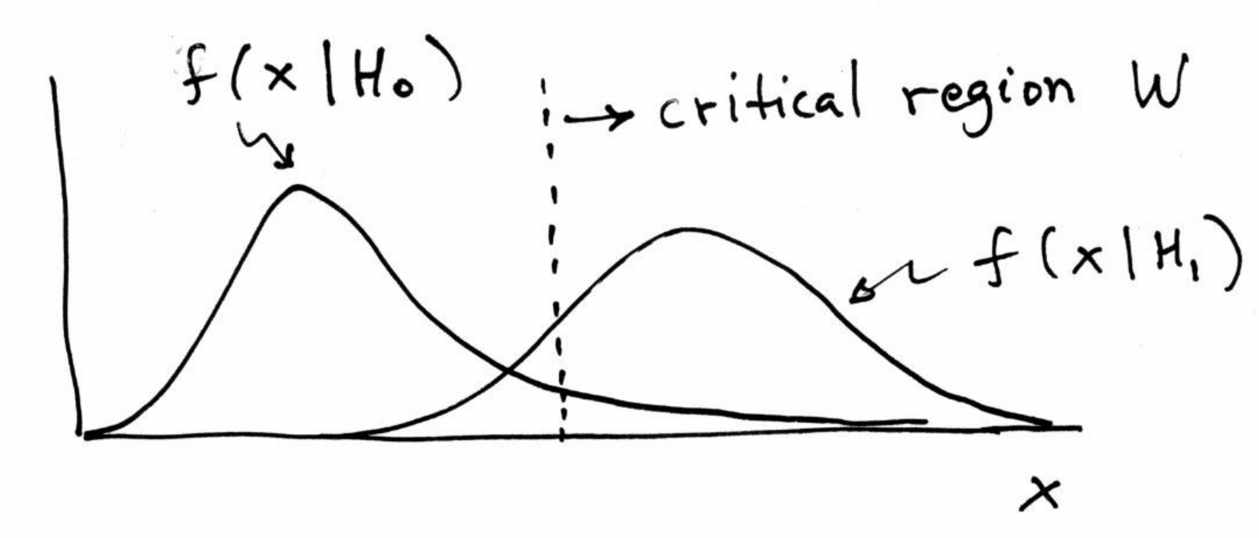
Critical region also called “rejection” region; complement is acceptance region.

Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)\,dH}$$

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in \mathrm{S} - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Physics context of a statistical test

Event Selection: the event types in question are both known to exist.

Example: separation of different particle types (electron vs muon) or known event types (ttbar vs QCD multijet).
Use the selected sample for further study.

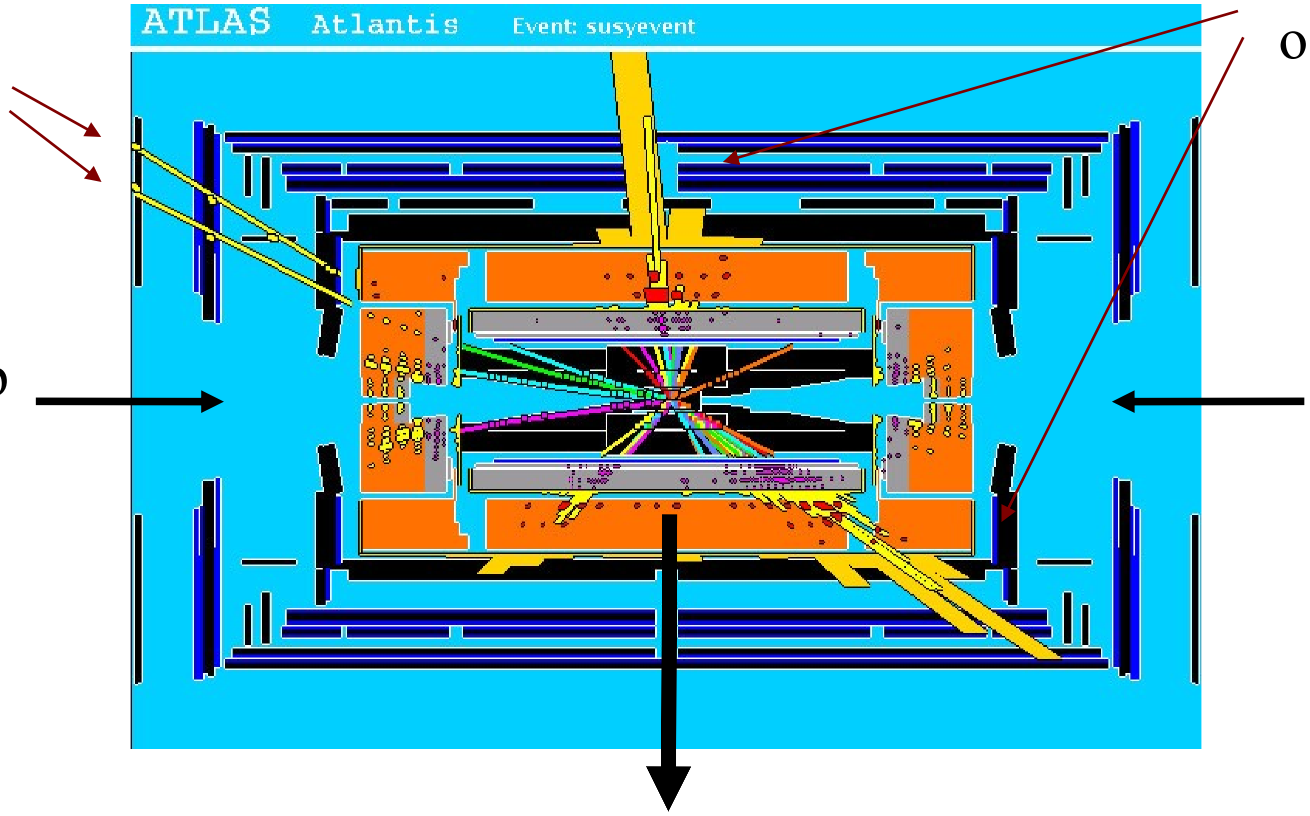
Search for New Physics: the null hypothesis H_0 means Standard Model events, and the alternative H_1 means "events of a type whose existence is not yet established" (to establish or exclude the signal model is the goal of the analysis).

Many subtle issues here, mainly related to the heavy burden of proof required to establish presence of a new phenomenon.

The optimal statistical test for a search is closely related to that used for event selection.

Suppose we want to discover this…

SUSY event (ATLAS simulation):



high p_T muons

high p_T jets of hadrons

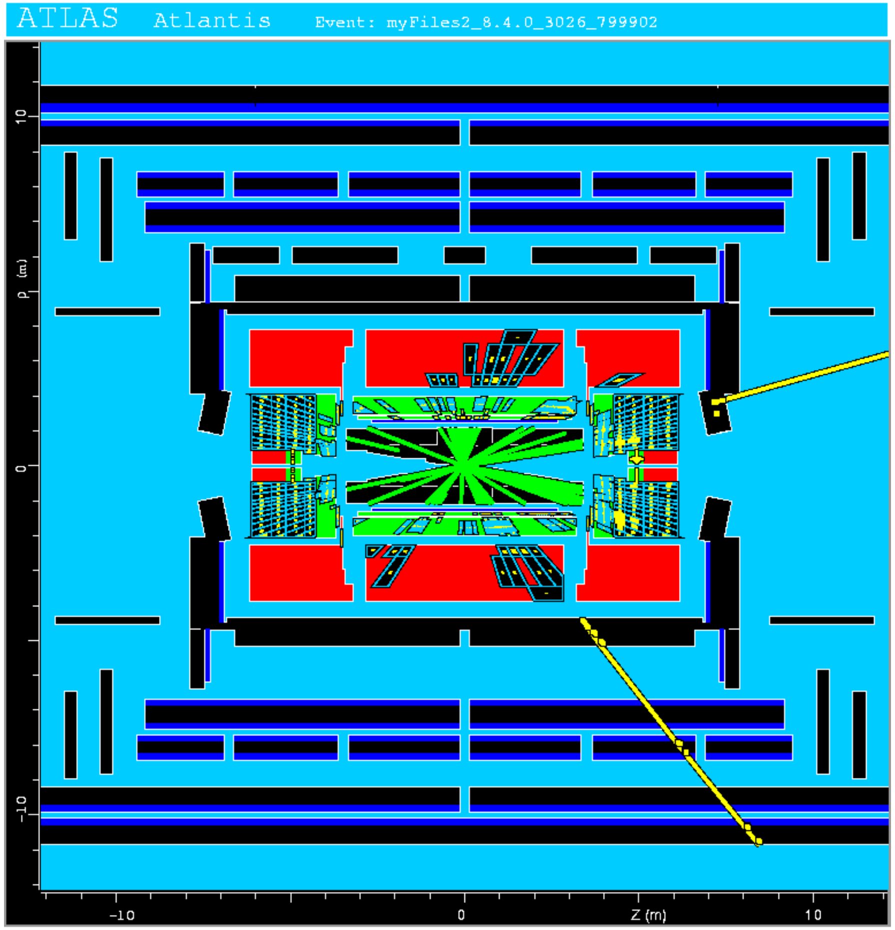
p

p

missing transverse energy

But we know we'll have lots of this…

ttbar event (ATLAS simulation)



SM event also has high p_T jets and muons, and missing transverse energy.

→ can easily mimic a SUSY event and thus constitutes a background.

Example of a multivariate statistical test

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

$$\mathrm{pp} \to t\bar{t}\ , \qquad \mathrm{pp} \to \tilde{g}\tilde{g}\ , \ldots$$

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.

Often call H_0 the background hypothesis (e.g. SM events); H_1, H_2, ... are possible signal hypotheses.

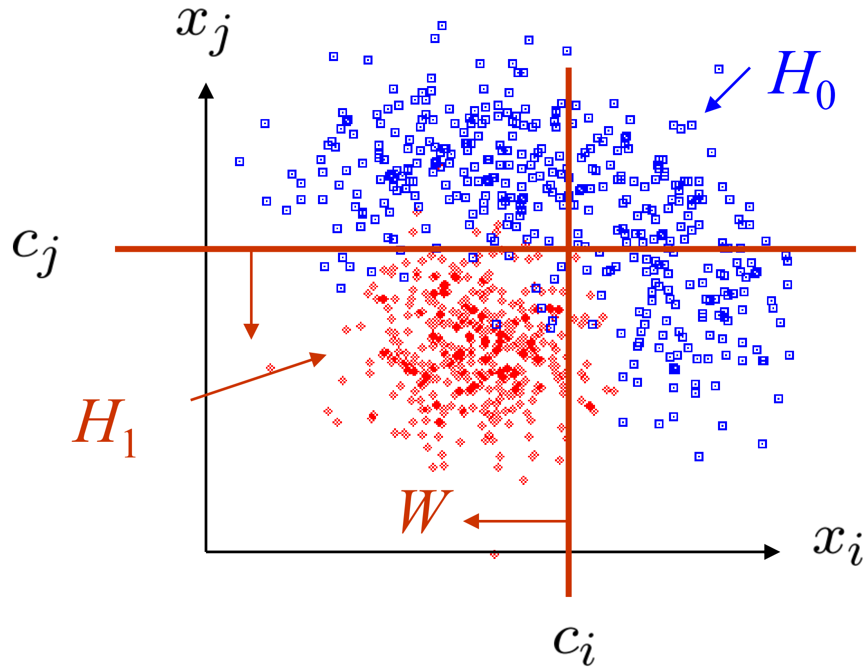
Defining a multivariate critical region

Each event is a point in ***x***-space; critical region is now defined by a ‘decision boundary’ in this space.

What is best way to determine the decision boundary?

Perhaps with ‘cuts’:

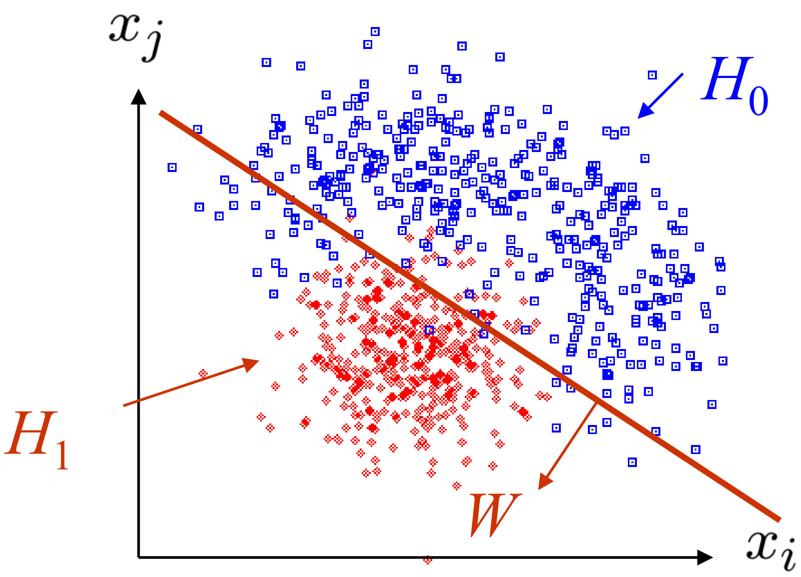
$$x_i < c_i$$
$$x_j < c_j$$



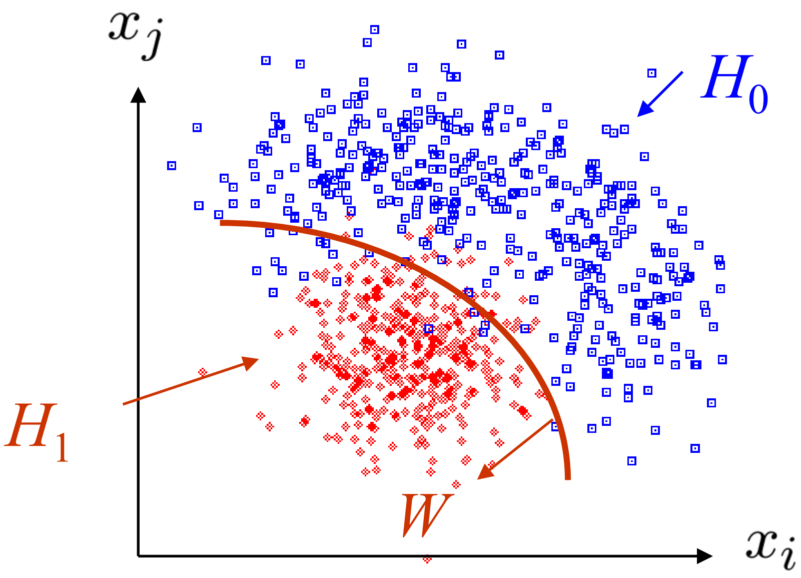
Other multivariate decision boundaries

Or maybe use some other sort of decision boundary:

linear



or nonlinear



Test statistics

The decision boundary can be defined by an equation of the form

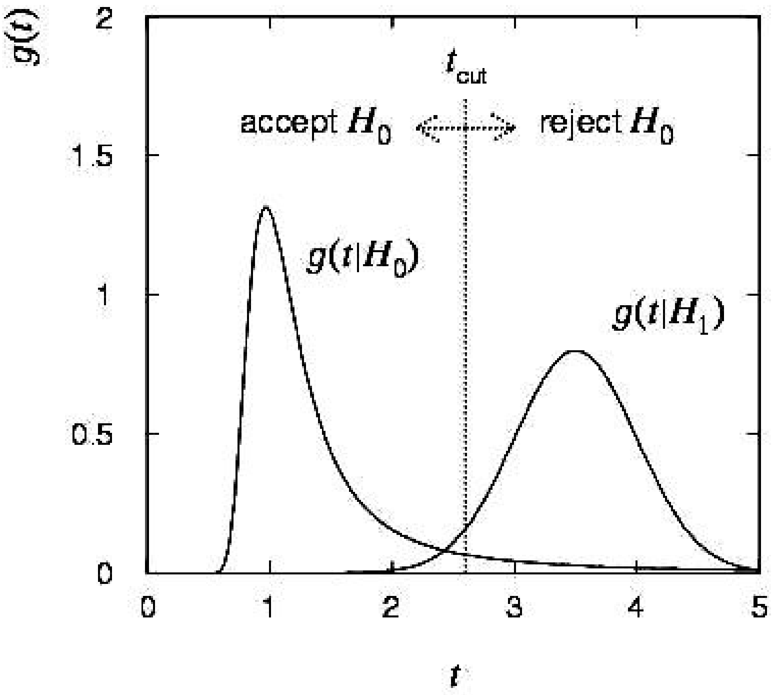
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Constructing a test statistic

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test H_0, (background) versus H_1, (signal) (highest ε_s for a given ε_b) choose the critical (rejection) region such that

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

where c is a constant which determines the power.

Equivalently, optimal scalar test statistic is $\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$

N.B. any monotonic function of this is leads to the same test.

Neyman-Pearson doesn't always help

The problem is that we usually don't have explicit formulae for the pdfs $p(\boldsymbol{x}|\mathrm{s})$, $p(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the likelihood ratio.

Instead we have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate $\vec{x} \sim p(\vec{x}|\mathrm{s}) \longrightarrow \vec{x}_1, \dots, \vec{x}_{N_s}$

generate $\vec{x} \sim p(\vec{x}|\mathrm{b}) \longrightarrow \vec{x}_1, \dots, \vec{x}_{N_b}$

"training data"

events of known type

Naive try: enter each (s,b) event into an n-dimensional histogram, use e.g. M bins for each of the n dimensions, total of M^n cells.

n is potentially large → prohibitively large number of cells to populate, can't generate enough training data.

Multivariate methods

Many new (and some old) methods for finding decision boundary:

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

New software for HEP, e.g.,

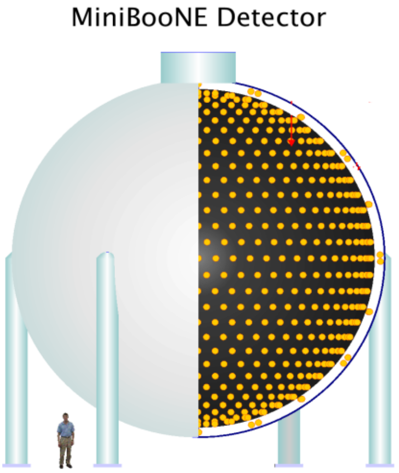
TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

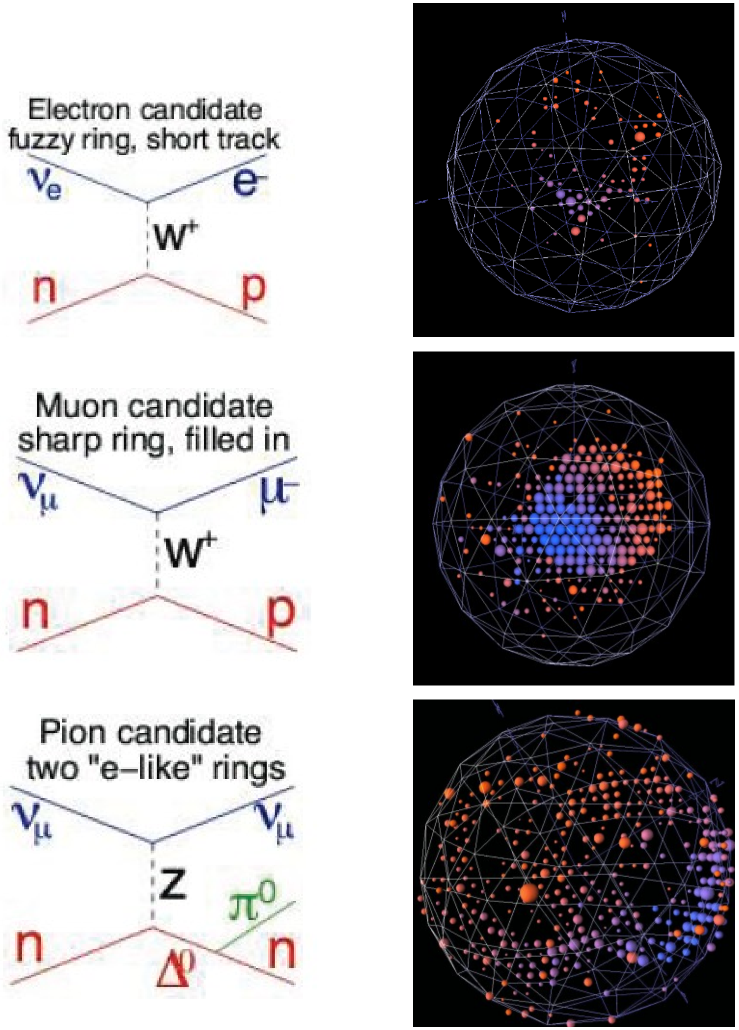
For more see e.g. references at end of this lecture.

For the rest of these lectures, I will focus on other aspects of tests, e.g., discovery significance and exclusion limits.

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:





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

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

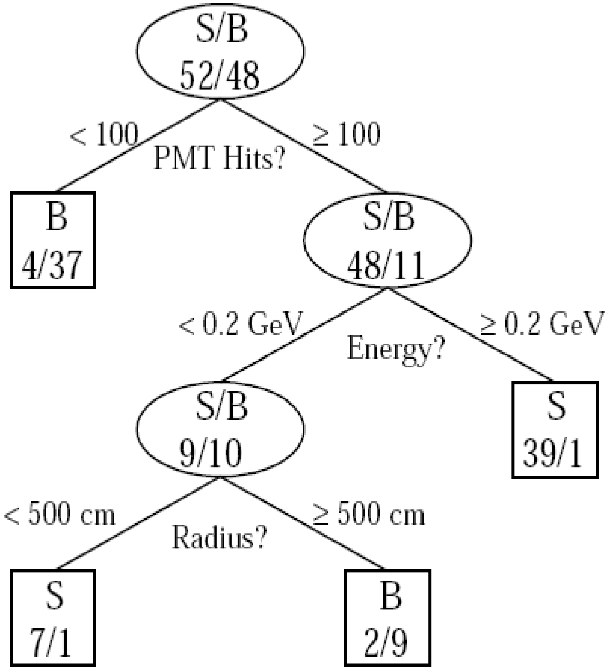
$$P = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}$$

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 the decision boundary.



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

Decision trees (2)

The terminal nodes (leaves) are classified as signal or background depending on majority vote (or e.g. signal fraction greater than a specified threshold).

This classifies every point in input-variable space as either signal or background, a decision tree classifier, with the discriminant function

$$f(\mathbf{x}) = 1 \text{ if } \mathbf{x} \in \text{signal region}, -1 \text{ otherwise}$$

Decision trees tend to be very sensitive to statistical fluctuations in the training sample.

Methods such as boosting can be used to stabilize the tree.

Boosting

Boosting is a general method of creating a set of classifiers which can be combined to achieve a new classifier that is more stable and has a smaller error than any individual one.

Often applied to decision trees but, can be applied to any classifier.

Suppose we have a training sample T consisting of N events with

$\boldsymbol{x}_1, \ldots., \boldsymbol{x}_N$	event data vectors (each $\boldsymbol{x}$ multivariate)
$y_1, \ldots., y_N$	true class labels, +1 for signal, -1 for background
$w_1, \ldots., w_N$	event weights

Now define a rule to create from this an ensemble of training samples T_1, T_2, ... , derive a classifier from each and average them.

AdaBoost

A successful boosting algorithm is AdaBoost (Freund & Schapire, 1997).

First initialize the training sample T_1 using the original

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ event data vectors

$y_1, \ldots, y_N$ true class labels (+1 or -1)

$w_1^{(1)}, \ldots, w_N^{(1)}$ event weights

with the weights equal and normalized such that $\sum_{i=1}^{N} w_i^{(1)} = 1$.

Train the classifier $f_1(\boldsymbol{x})$ (e.g. a decision tree) using the weights $\boldsymbol{w}^{(1)}$

so as to minimize the classification error rate,

$$\varepsilon_1 = \sum_{i=1}^{N} w_i^{(1)} I(y_i f_1(\mathbf{x}_i) \leqslant 0) ,$$

where $I(X) = 1$ if X is true and is zero otherwise.

Updating the event weights (AdaBoost)

Assign a score to the *k*th classifier based on its error rate:

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

Define the training sample for step *k*+1 from that of *k* by updating the event weights according to

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

i = event index

k = training sample index

Normalize so that

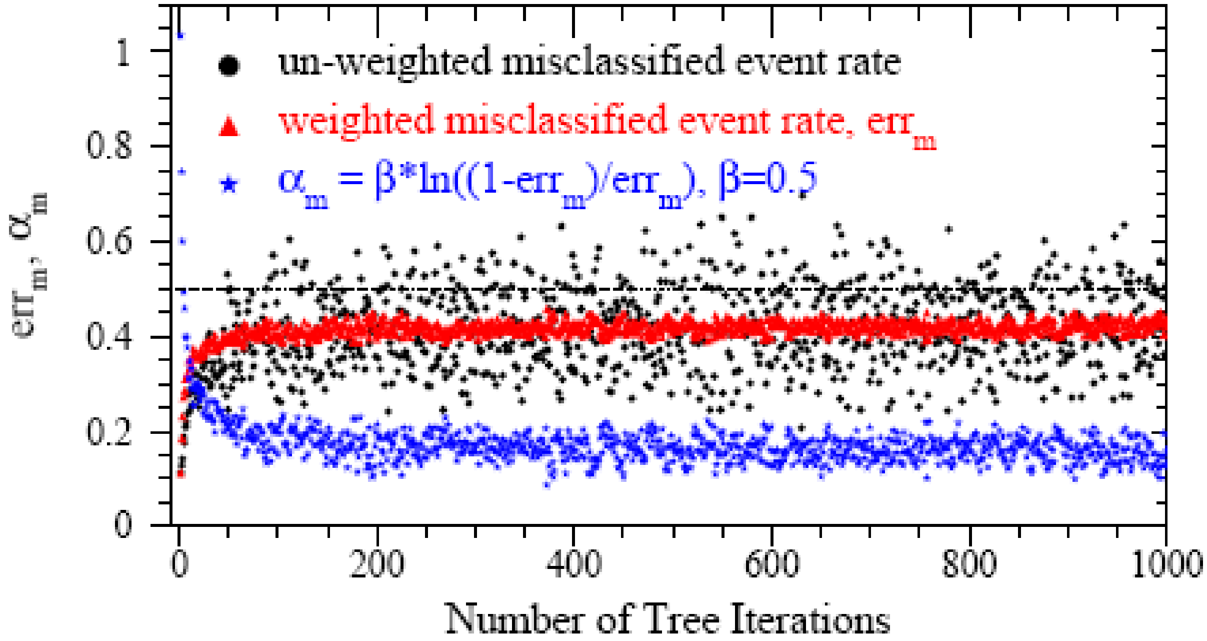
$$\sum_i w_i^{(k+1)} = 1$$

Iterate *K* times, final classifier is $y(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{x}, T_k)$

BDT example from MiniBooNE

~200 input variables for each event (ν interaction producing e, μ or π).

Each individual tree is relatively weak, with a misclassification error rate ~ 0.4 – 0.45

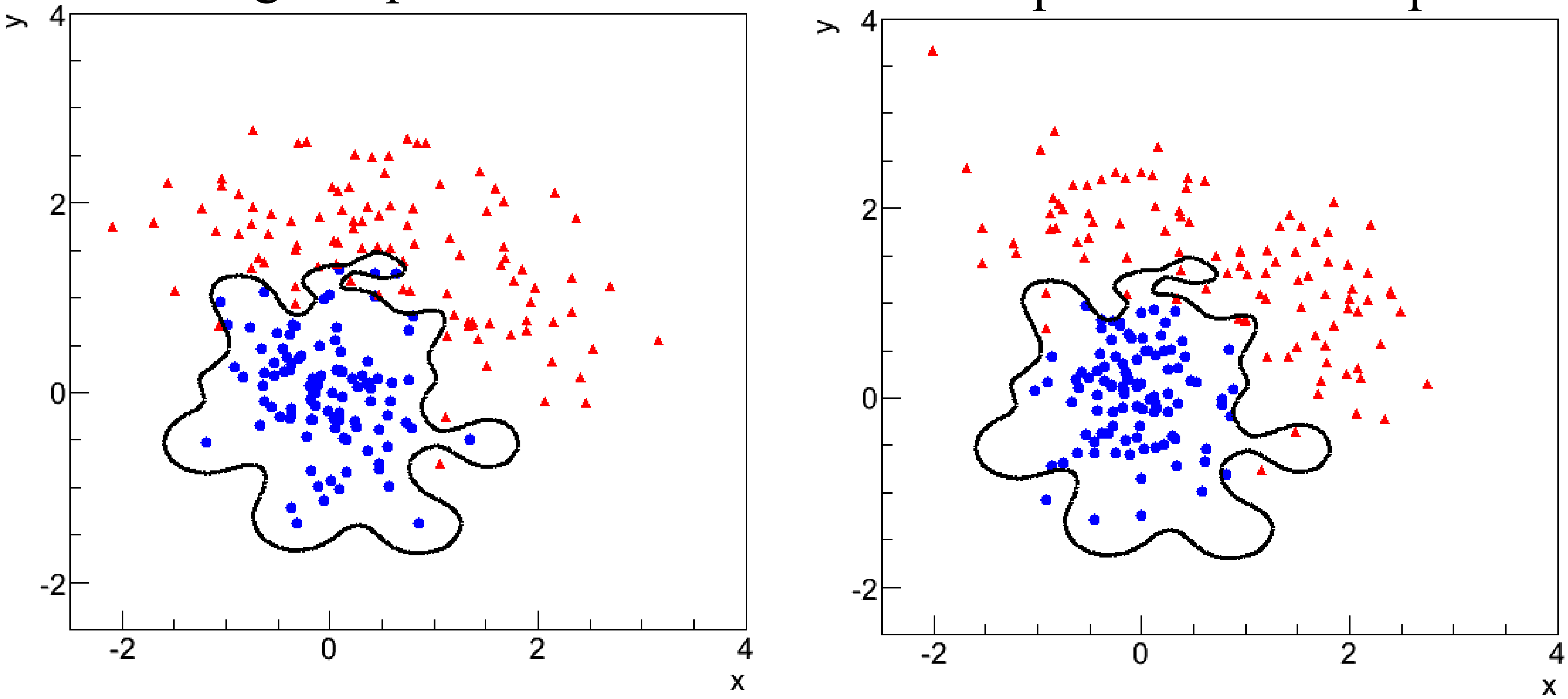


B. Roe et al., NIM 543 (2005) 577

Overtraining

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

Monitor by applying classifier to independent test sample.



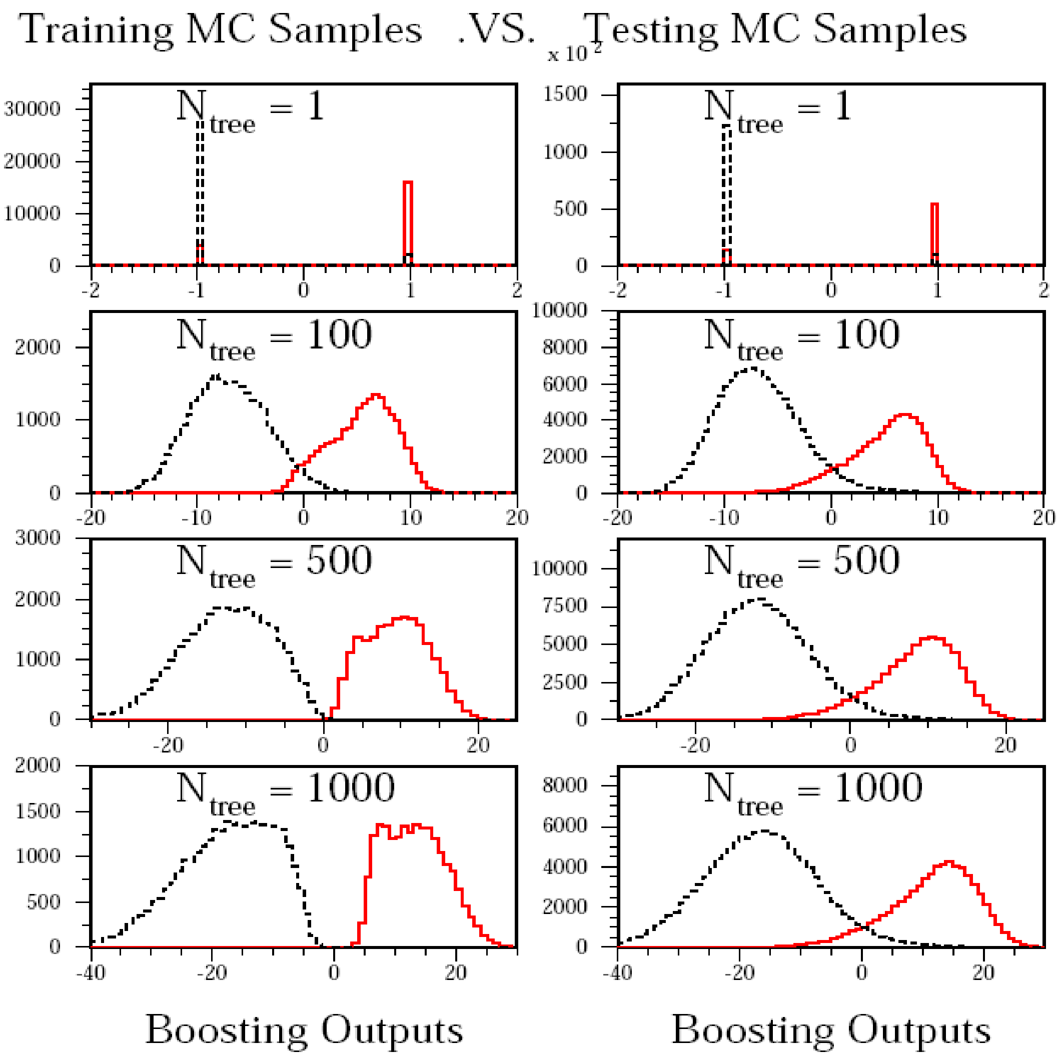
training sample

independent test sample

Monitoring overtraining

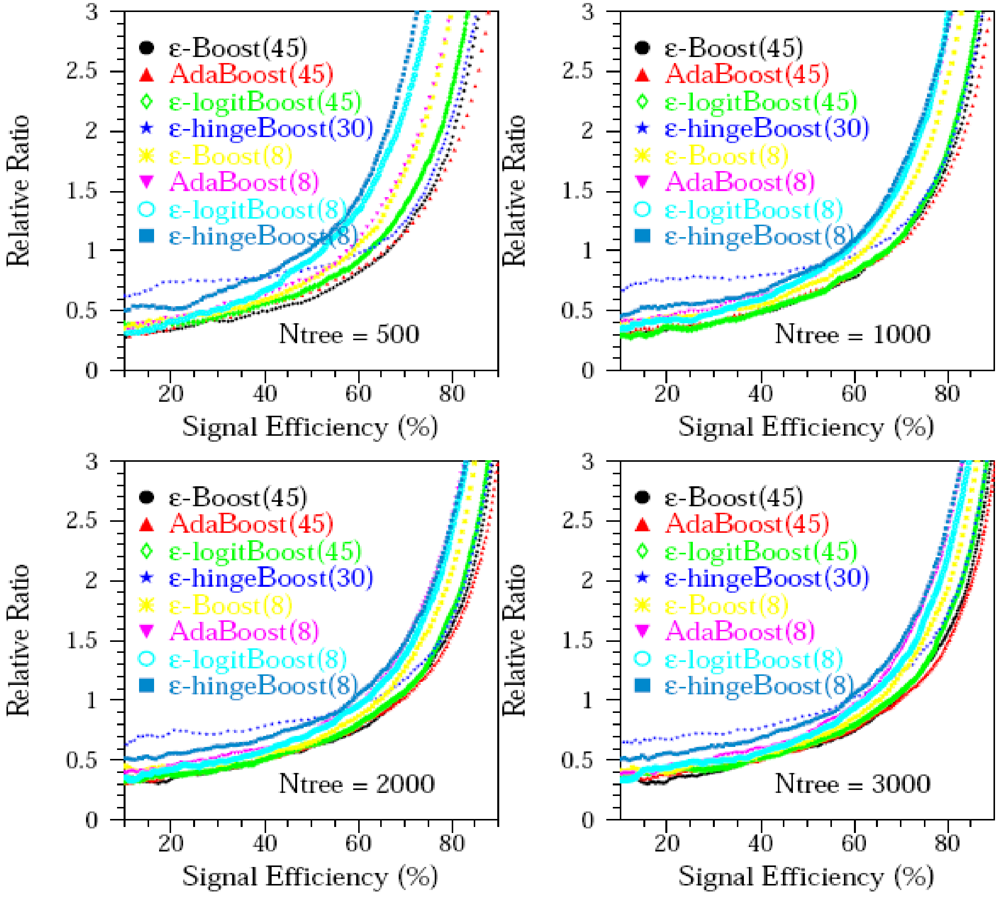
From MiniBooNE example:

Performance stable after a few hundred trees.



Comparison of boosting algorithms

A number of boosting algorithms on the market; differ in the update rule for the weights.



Boosted decision tree summary

Advantage of boosted decision tree is it can handle a large number of inputs. Those that provide little/no separation are rarely used as tree splitters are effectively ignored.

Easy to deal with inputs of mixed types (real, integer, categorical...).

If a tree has only a few leaves it is easy to visualize (but rarely use only a single tree).

There are a number of boosting algorithms, which differ primarily in the rule for updating the weights (ε-Boost, LogitBoost,...)

Other ways of combining weaker classifiers: Bagging (Boostrap-Aggregating), generates the ensemble of classifiers by random sampling with replacement from the full training sample.

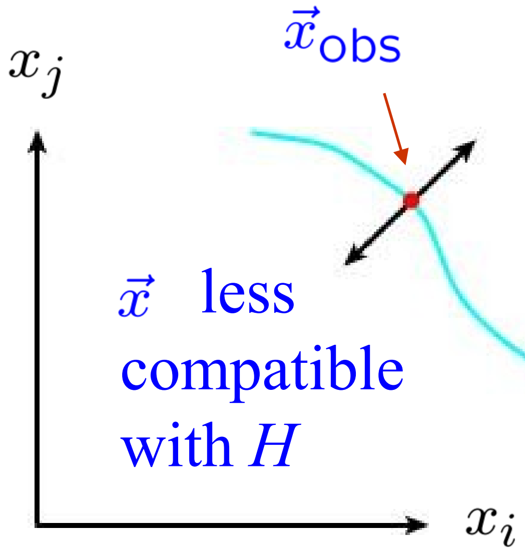
Testing significance / goodness-of-fit

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of *H* in light of the data?

Decide what part of the data space represents less compatibility with *H* than does the point $\vec{x}_{\mathrm{obs}}$. (Not unique!)



$\vec{x}$ more compatible with *H*

$\vec{x}$ less compatible with *H*

p-values

Express level of agreement between data and *H* with *p*-value:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

⚠ This is not the probability that *H* is true!

In frequentist statistics we don't talk about *P*(*H*) (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

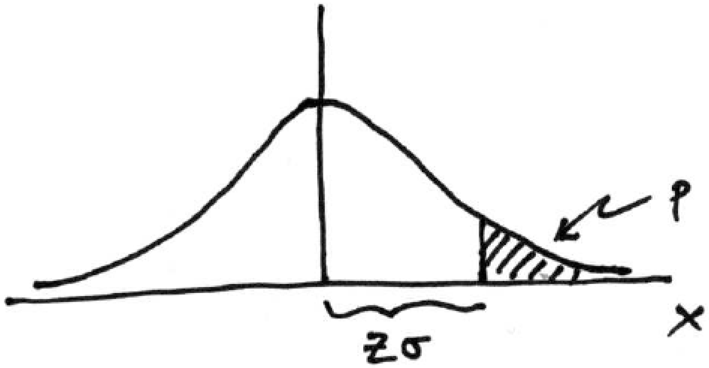
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as *P*(*H*).

Significance from *p*-value

Often define significance *Z* as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same *p*-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$

`1 - TMath::Freq`

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

`TMath::NormQuantile`

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

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

Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
D^0D^0 mixing	~0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

One should also consider the degree to which the data are compatible with the new phenomenon and possible systematic errors in the model on which the p-value is based: p-value is only first step!

Distribution of the *p*-value

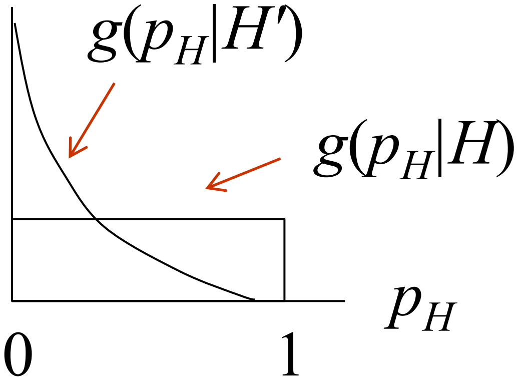
The *p*-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the *p*-value of *H* is found from a test statistic $t(\boldsymbol{x})$ as

$$p_H = \int_t^\infty f(t'|H)\,dt'$$

The pdf of p_H under assumption of *H* is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

In general for continuous data, under assumption of *H*, $p_H \sim$ Uniform[0,1] and is concentrated toward zero for Some (broad) class of alternatives.



Using a *p*-value to define test of H_0

So the probability to find the *p*-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We started by defining critical region in the original data space (***x***), then reformulated this in terms of a scalar test statistic t(***x***).

We can take this one step further and define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the *p*-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ .

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size α (confidence level is $1 - \alpha$).

The interval will cover the true value of θ with probability $\geq 1 - \alpha$.

Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

Relation between confidence interval and *p*-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a *p*-value, p_θ..

If $p_\theta < \alpha$, then we reject θ.

The confidence interval at CL = 1 – α consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta \geq \alpha$.

In practice find by setting $p_\theta = \alpha$ and solve for θ.

A simple example

For each event we measure two variables, $\boldsymbol{x} = (x_1, x_2)$.

Suppose that for background events (hypothesis H_0),

$$f(\mathbf{x}|H_0) = \frac{1}{\xi_1} e^{-x_1/\xi_1} \, \frac{1}{\xi_2} e^{-x_2/\xi_2}$$

and for a certain signal model (hypothesis H_1) they follow

$$f(\mathbf{x}|H_1) = C \, \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \, \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x_2-\mu_2)^2/2\sigma_2^2}$$

where $x_1, x_2 \geq 0$ and C is a normalization constant.

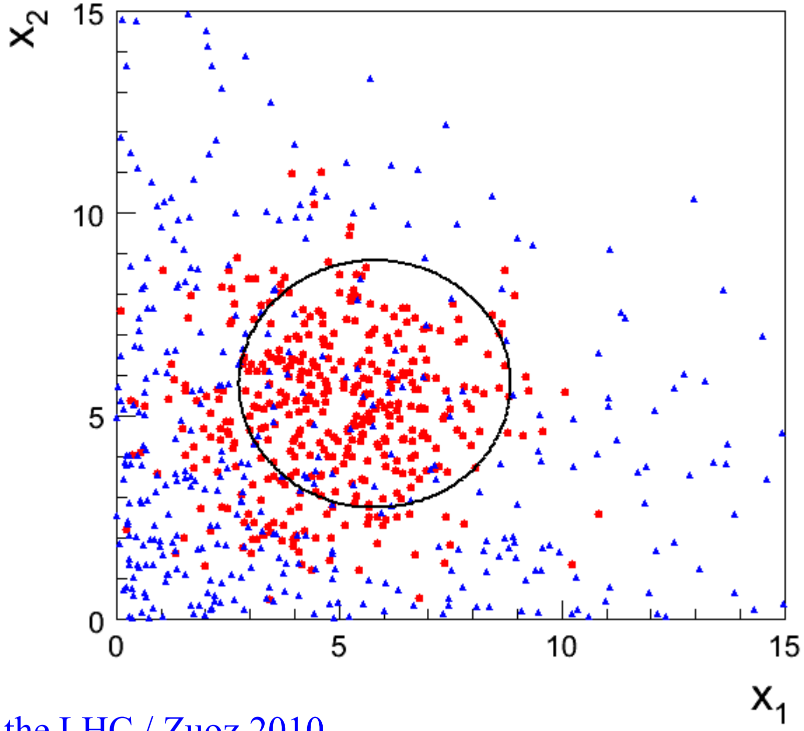
Likelihood ratio as test statistic

In a real-world problem we usually wouldn't have the pdfs $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$, so we wouldn't be able to evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

for a given observed $\boldsymbol{x}$, hence the need for multivariate methods to approximate this with some other function.

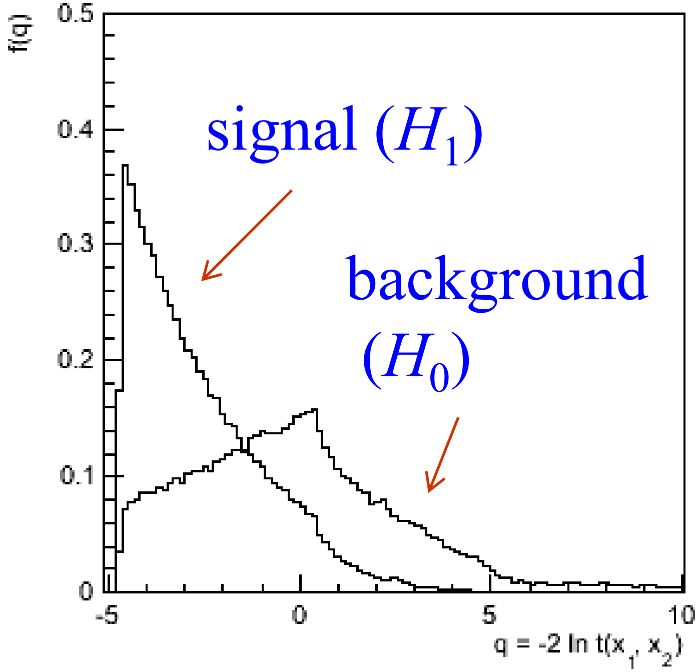
But in this example we can find contours of constant likelihood ratio such as:



Event selection using the LR

Using Monte Carlo, we can find the distribution of the likelihood ratio or equivalently of

$$q = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - \frac{2x_1}{\xi_1} - \frac{2x_2}{\xi_2} = -2\ln t(\mathbf{x}) + C$$



From the Neyman-Pearson lemma we know that by cutting on this variable we would select a signal sample with the highest signal efficiency (test power) for a given background efficiency.

Search for the signal process

But what if the signal process is not known to exist and we want to search for it. The relevant hypotheses are therefore

H_0: all events are of the background type
H_1: the events are a mixture of signal and background

Rejecting H_0 with $Z > 5$ constitutes "discovering" new physics.

Suppose that for a given integrated luminosity, the expected number of signal events is s, and for background b.

The observed number of events n will follow a Poisson distribution:

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

Likelihoods for full experiment

We observe *n* events, and thus measure *n* instances of $\boldsymbol{x} = (x_1, x_2)$.

The likelihood function for the entire experiment assuming the background-only hypothesis (H_0) is

$$L_b = \frac{b^n}{n!} e^{-b} \prod_{i=1}^{n} f(\mathbf{x}_i|\mathrm{b})$$

and for the "signal plus background" hypothesis (H_1) it is

$$L_{s+b} = \frac{(s+b)^n}{n!} e^{-(s+b)} \prod_{i=1}^{n} \left(\pi_{\mathrm{s}} f(\mathbf{x}_i|\mathrm{s}) + \pi_{\mathrm{b}} f(\mathbf{x}_i|\mathrm{b})\right)$$

where π_s and π_b are the (prior) probabilities for an event to be signal or background, respectively.

Likelihood ratio for full experiment

We can define a test statistic *Q* monotonic in the likelihood ratio as

$$Q = -2\ln\frac{L_{s+b}}{L_b} = -s + \sum_{i=1}^{n} \ln\left(1 + \frac{s}{b}\frac{f(\mathbf{x}_i|\mathrm{s})}{f(\mathbf{x}_i|\mathrm{b})}\right)$$

To compute *p*-values for the b and s+b hypotheses given an observed value of *Q* we need the distributions *f*(*Q*|b) and *f*(*Q*|s+b).

Note that the term –*s* in front is a constant and can be dropped.

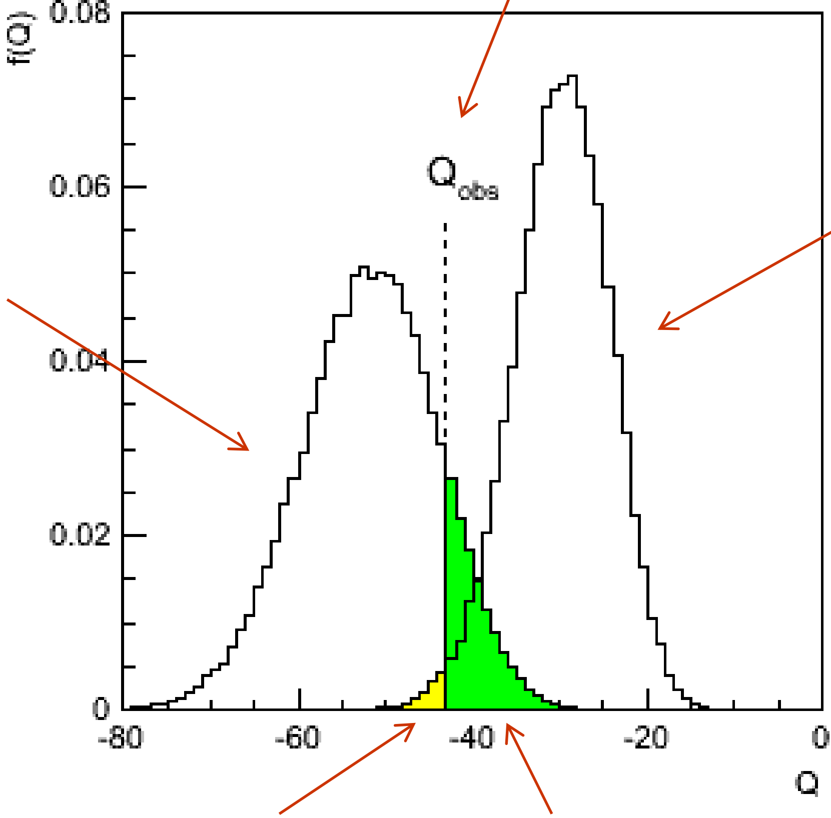
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

Can exploit this to relate distribution of *Q* to that of single event terms using (Fast) Fourier Transforms (Hu and Nielsen, physics/9906010).

Distribution of *Q*

Take e.g. b = 100, s = 20.

Suppose in real experiment *Q* is observed here.



f (*Q*|s+b)

f (*Q*|b)

p-value of b only

p-value of s+b

Systematic uncertainties

Up to now we assumed all parameters were known exactly.

In practice they have some (systematic) uncertainty.

Suppose e.g. uncertainty in expected number of background events b is characterized by a (Bayesian) pdf $\pi(b)$.

Maybe take a Gaussian, i.e.,

$$\pi(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_0)^2/2\sigma_b^2}$$

where b_0 is the nominal (measured) value and σ_b is the estimated uncertainty.

In fact for many systematics a Gaussian pdf is hard to defend – more on this later.

Distribution of Q with systematics

To get the desired p-values we need the pdf $f(Q)$, but this depends on b, which we don't know exactly.

But we can obtain the Bayesian model average:

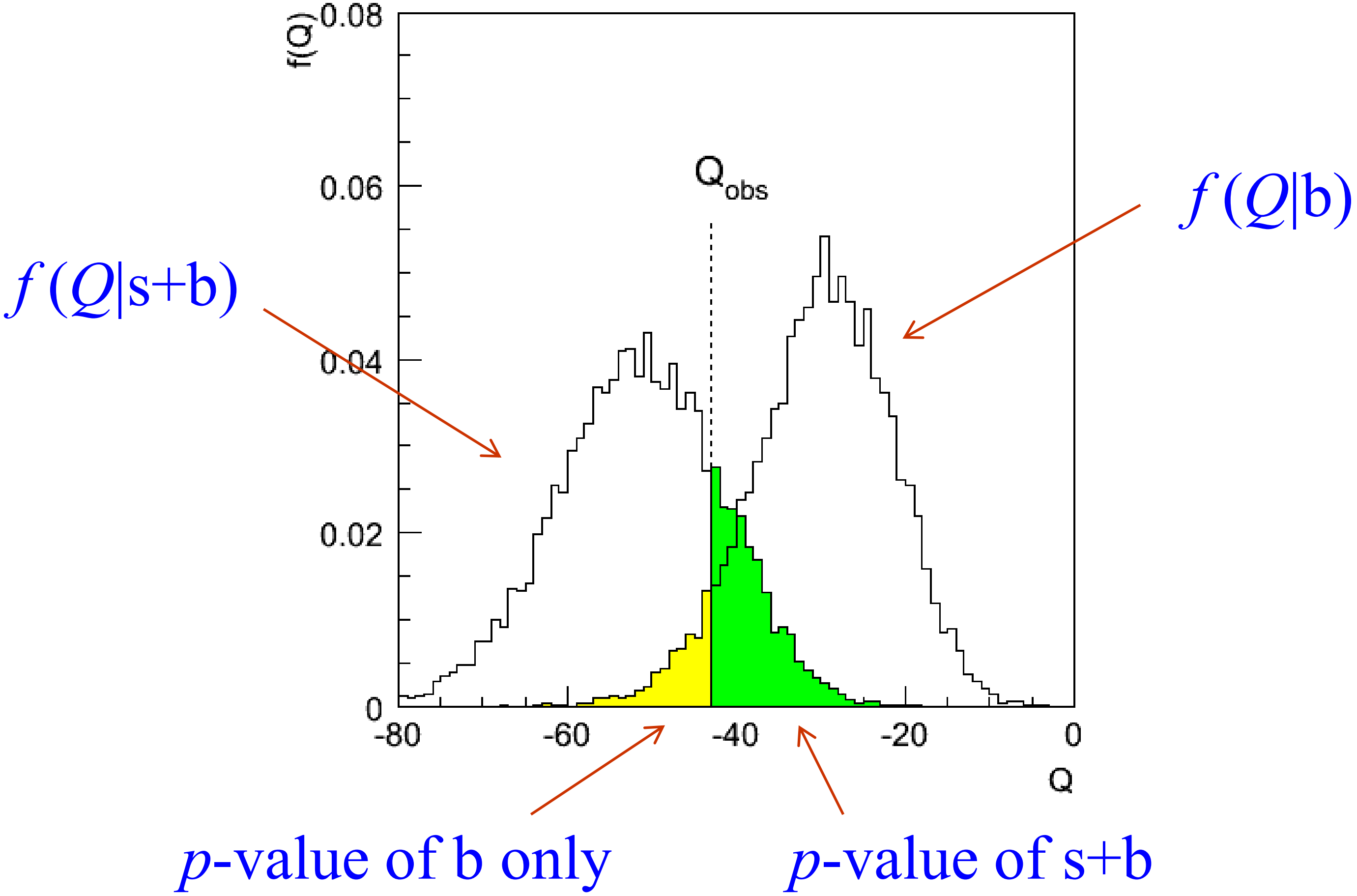
$$f(Q) = \int f(Q|b) \pi(b) \, db$$

With Monte Carlo, sample b from $\pi(b)$, then use this to generate Q from $f(Q|b)$, i.e., a new value of b is used to generate the data for every simulation of the experiment.

This broadens the distributions of Q and thus increases the p-value (decreases significance Z) for a given Q_{obs}.

Distribution of Q with systematics (2)

For $s = 20$, $b_0 = 100$, $\sigma_b = 10$ this gives



$f(Q|\mathrm{s+b})$

$f(Q|\mathrm{b})$

p-value of b only

p-value of s+b

Using the likelihood ratio $L(s)/L(\hat{s})$

Instead of the likelihood ratio L_{s+b}/L_b, suppose we use as a test statistic

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

maximizes $L(s)$

Intuitively this is a good measure of the level of agreement between the data and the hypothesized value of s.

low λ: poor agreement
high λ : good agreement
$0 \leq \lambda \leq 1$

$L(s)/L(\hat{s})$ for counting experiment

Consider an experiment where we only count n events with $n \sim \text{Poisson}(s + b)$. Then $\hat{s} = n - b$.

To establish discovery of signal we test the hypothesis $s = 0$ using

$$\ln \lambda(0) = n \ln(b) - b - n \ln n + n$$

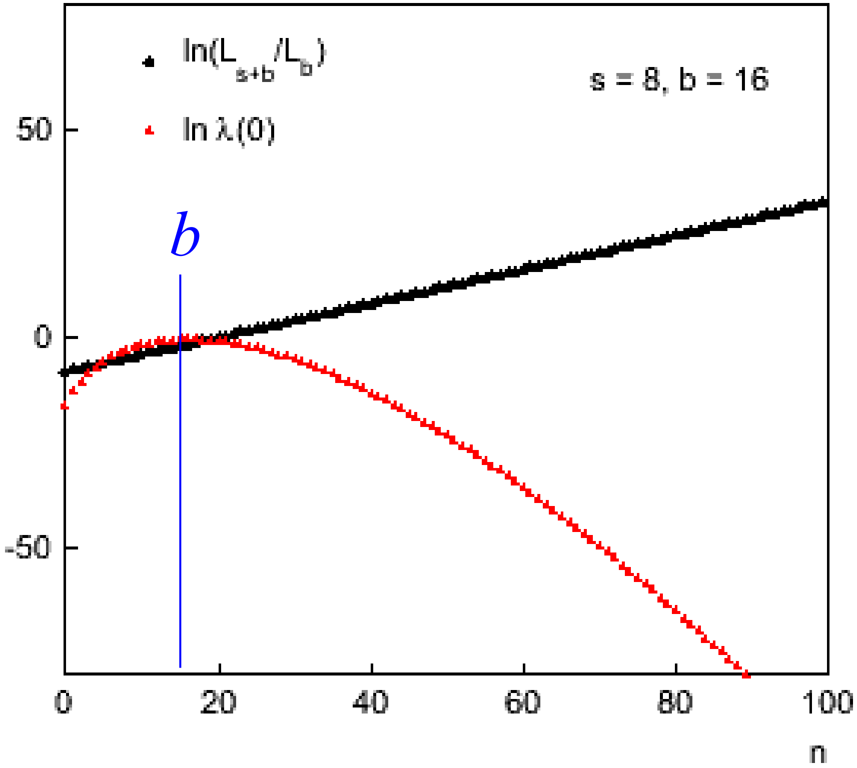
whereas previously we had used

$$\ln \frac{L_{s+b}}{L_b} = n \ln \left(1 + \frac{s}{b}\right) - s$$

which is monotonic in n and thus equivalent to using n as the test statistic.

$L(s)/L(\hat{s})$ for counting experiment (2)

But if we only consider the possibility of signal being present when $n > b$, then in this range $\lambda(0)$ is also monotonic in n, so both likelihood ratios lead to the same test.



$L(s)/L(\hat{s})$ for general experiment

If we do not simply count events but also measure for each some set of numbers, then the two likelihood ratios do not necessarily give equivalent tests, but in practice will be very close.

$\lambda(s)$ has the important advantage that for a sufficiently large event sample, its distribution approaches a well defined form (Wilks' Theorem).

In practice the approach to the asymptotic form is rapid and one obtains a good approximation even for relatively small data samples (but need to check with MC).

This remains true even when we have adjustable nuisance parameters in the problem, i.e., parameters that are needed for a correct description of the data but are otherwise not of interest (key to dealing with systematic uncertainties).

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

Prototype LHC 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

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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

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

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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 $\boldsymbol{\theta}$.

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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.

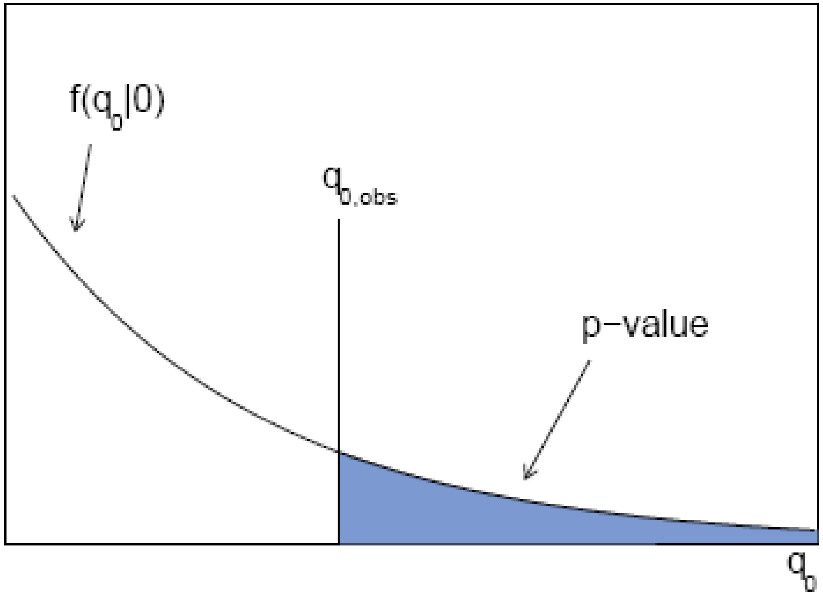
Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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

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

will get formula for this later



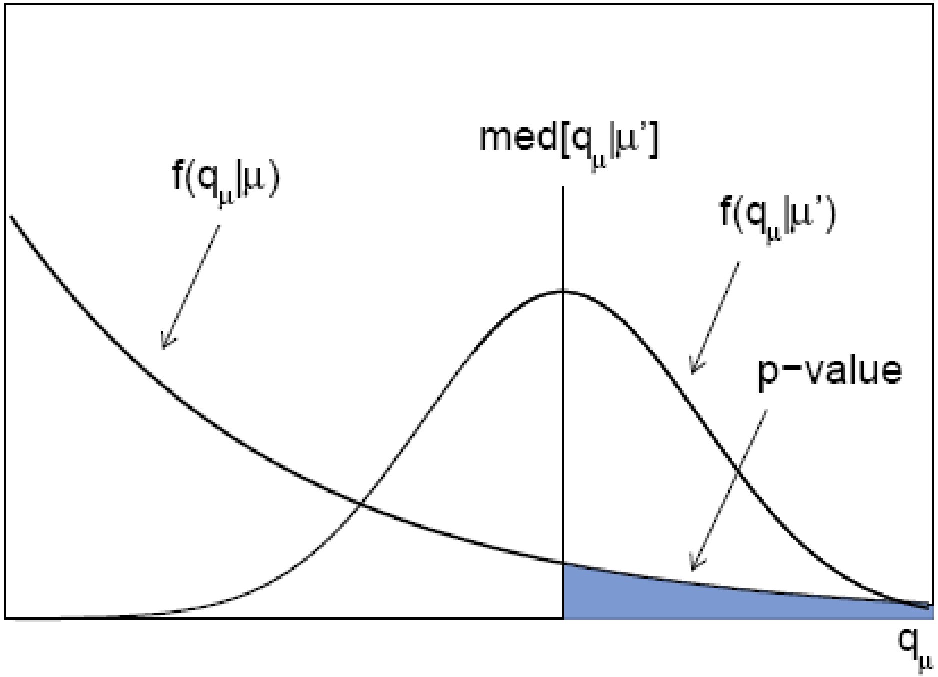
From *p*-value get equivalent significance,

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

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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 $f(q_0|\mu')$,

Distribution of q_0

Assuming the Wald approximation, 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:

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

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

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

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

Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

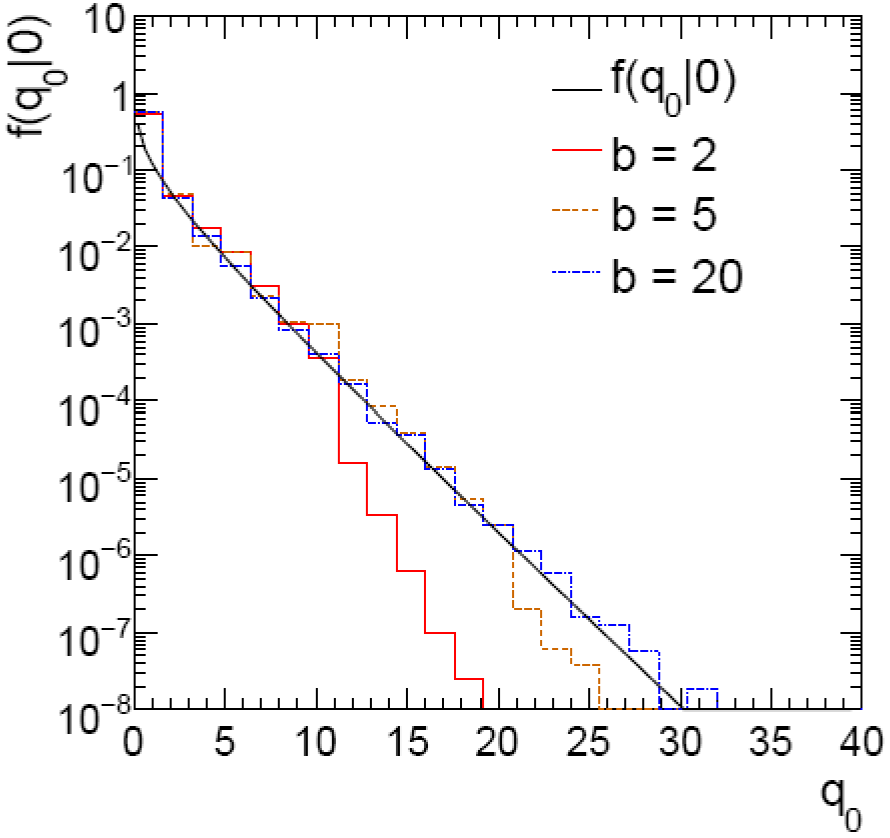
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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

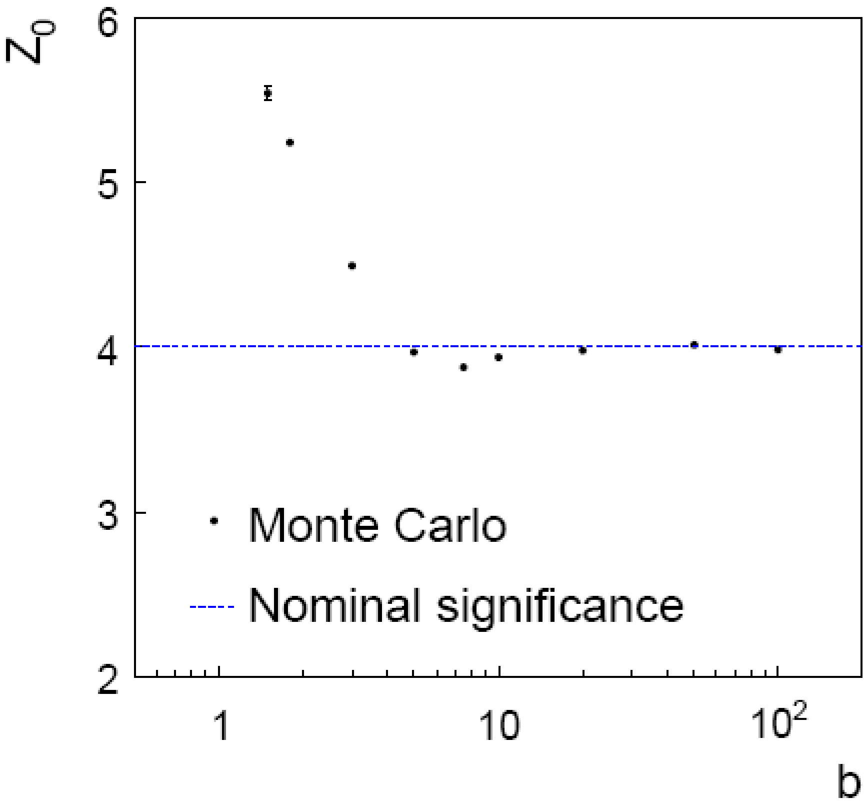


Monte Carlo test of asymptotic formulae

Significance from asymptotic formula, here $Z_0 = \sqrt{q_0} = 4$, compared to MC (true) value.

For very low *b*, asymptotic formula underestimates Z_0.

Then slight overshoot before rapidly converging to MC value.

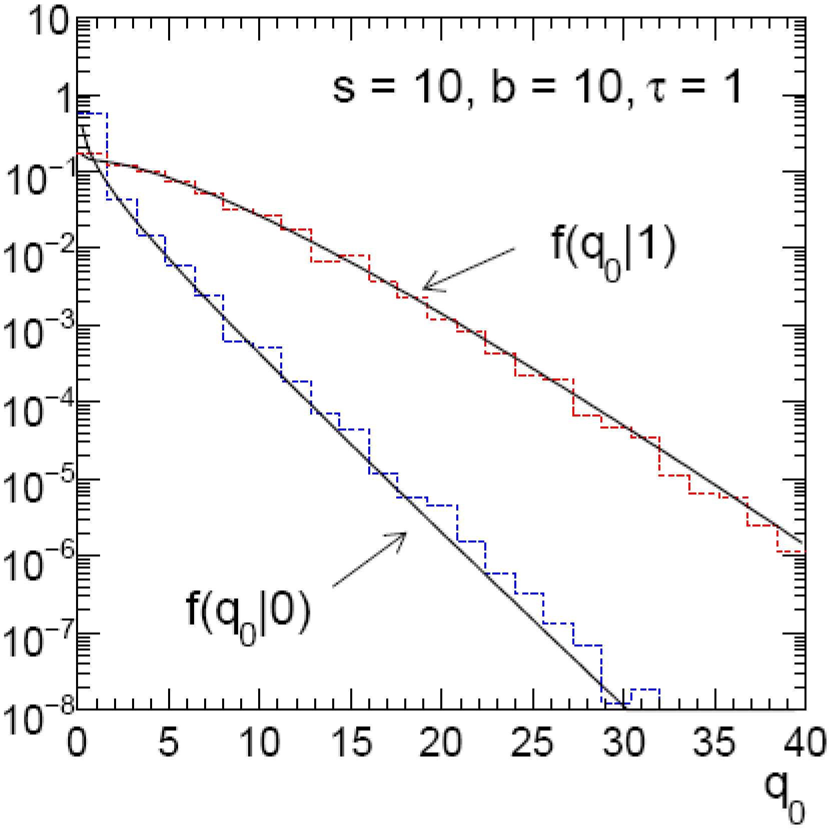


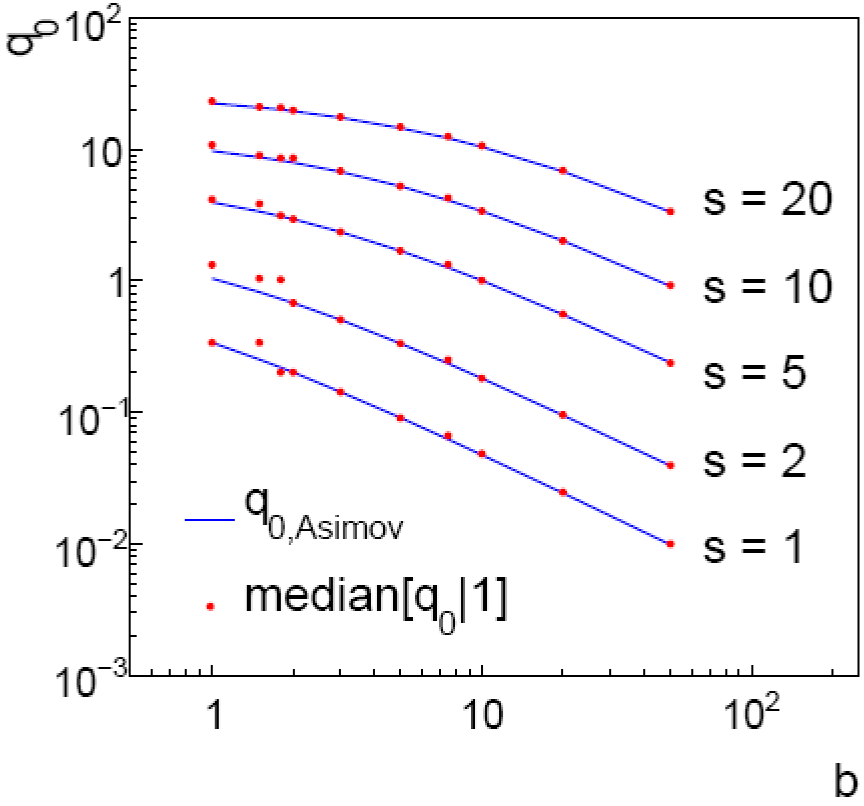
Cowan, Cranmer, Gross and Vitells, arXiv:1007.1727

Monte Carlo test of asymptotic formulae

Asymptotic $f(q_0|1)$ good already for fairly small samples.

Median[$q_0|1$] from Asimov data set; good agreement with MC.





Higgs search with profile likelihood

Combination of Higgs boson search channels (ATLAS)

Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics, arXiv:0901.0512, CERN-OPEN-2008-20.

Standard Model Higgs channels considered (more to be used later):

$H \rightarrow \gamma\gamma$

$H \rightarrow WW^{(*)} \rightarrow e\nu\mu\nu$

$H \rightarrow ZZ^{(*)} \rightarrow 4l \;\; (l = e, \mu)$

$H \rightarrow \tau^+\tau^- \rightarrow ll, lh$

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

An example: ATLAS Higgs search

(ATLAS Collab., CERN-OPEN-2008-020)

Statistical Combination of Several Important Standard Model Higgs Boson Search Channels.

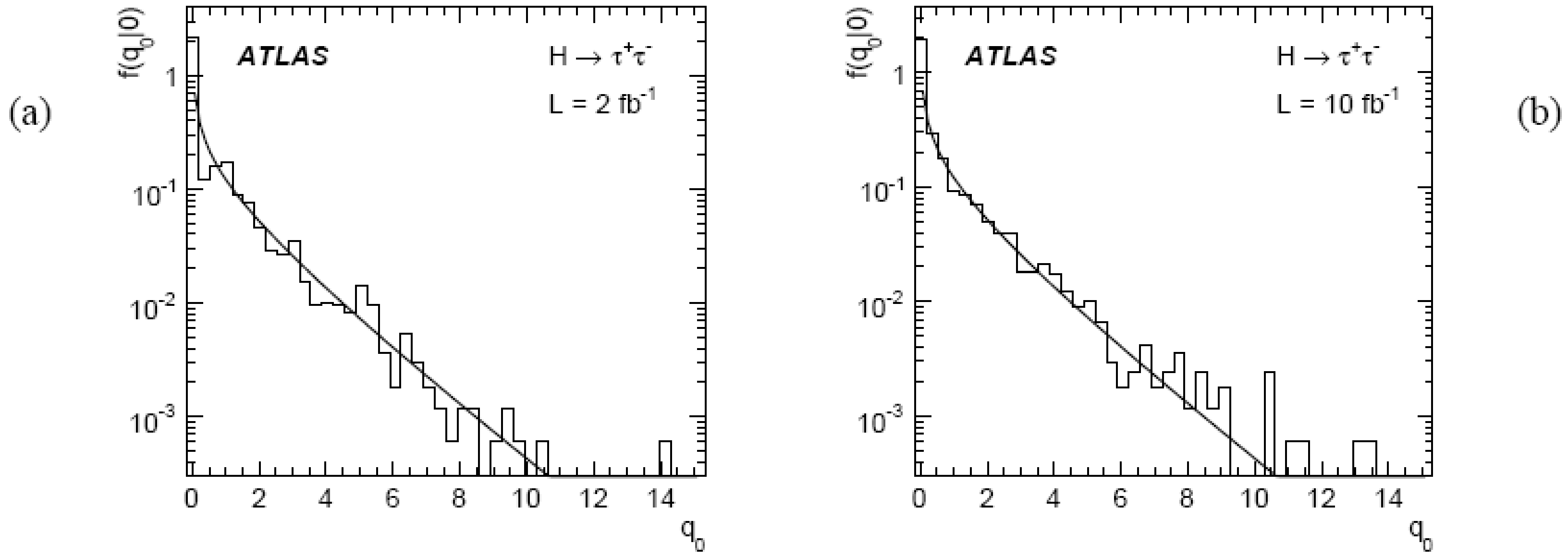
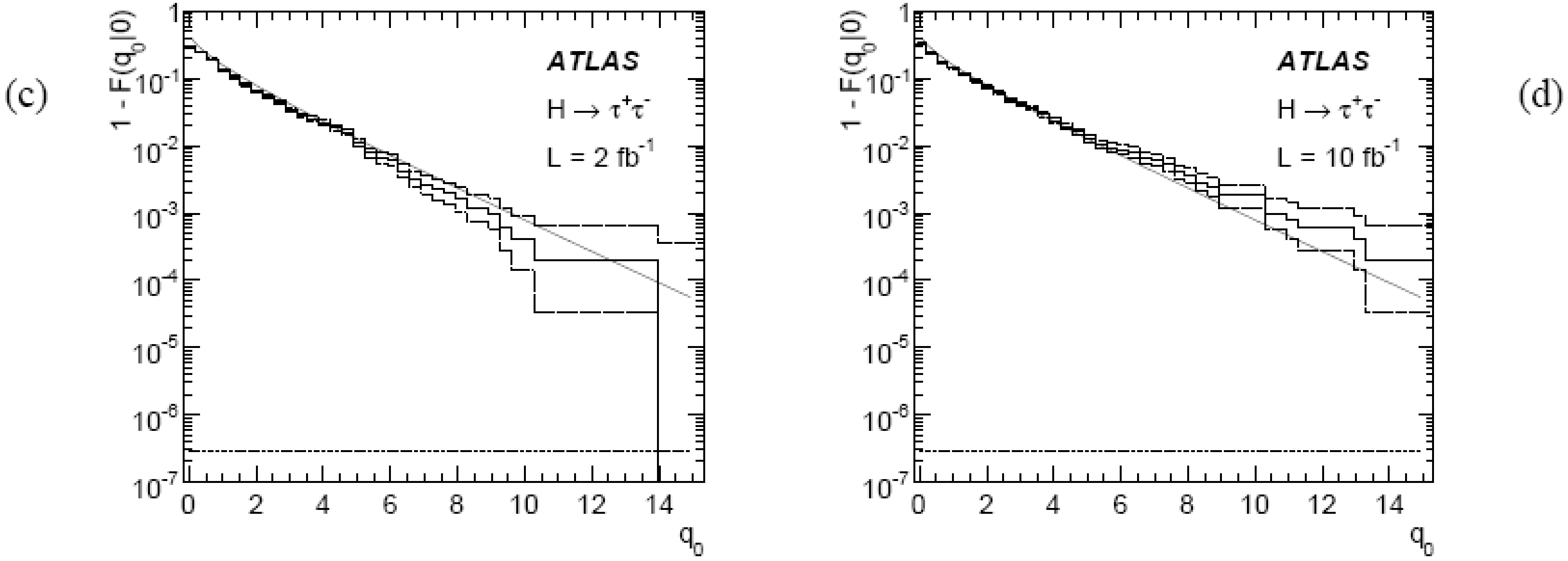


Figure 12: The distribution of the test statistic q_0 for $H \to \tau^+\tau^-$ under the null background-only hypothesis, for $m_H = 130\,\mathrm{GeV}$ with an integrated luminosity of 2 (a) and 10 (b) fb^{-1}. A $\frac{1}{2}\chi_1^2$ distribution is superimposed. Figures (c) and (d) show $1 - F(q_0)$ where $F(q_0)$ is the corresponding cumulative distribution. The small excess of events at high q_0 is statistically compatible with the expected curves, as can be seen by comparison with the dotted histograms that show the 68.3% central confidence intervals for $p = 1 - F(q_0|0)$. The lower dotted line at 2.87×10^{-7} shows the 5σ discovery threshold.

Cumulative distributions of q_0

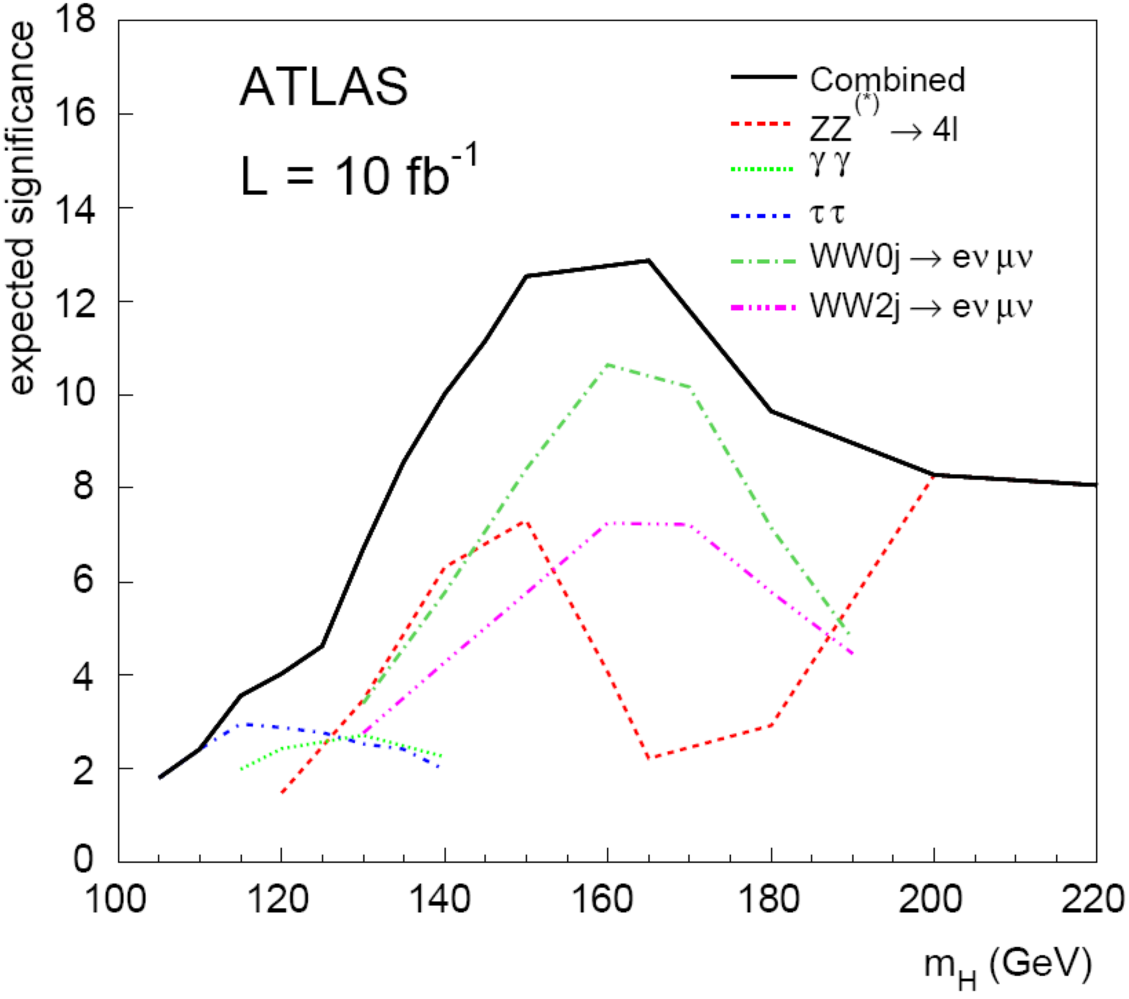
To validate to 5σ level, need distribution out to q_0 = 25, i.e., around 10^8 simulated experiments.

Will do this if we really see something like a discovery.



Combined median significance

ATLAS arXiv:0901.0512



N.B. illustrates statistical method, but study did not include all usable Higgs channels.

Profile likelihood ratio for upper limits

For purposes of setting an upper limit on μ use

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

Note for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ.

Note also here we allow the estimator for μ be negative (but $\hat{\mu} s_i + b_i$ must be positive).

Alternative test statistic for upper limits

Assume physical signal model has $\mu > 0$, therefore if estimator for μ comes out negative, the closest physical model has $\mu = 0$.

Therefore could also measure level of discrepancy between data and hypothesized μ with

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\[2ex] \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0\,. \end{cases} \qquad \tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

Performance not identical to but very close to q_μ (of previous slide).

q_μ is simpler in important ways.

Distribution of q_μ

Similar results for q_μ

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

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

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

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

Distribution of $\tilde{q}_\mu$

Similar results for $\tilde{q}_\mu$

$$
\begin{aligned}
f(\tilde{q}_\mu|\mu') &= \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(\tilde{q}_\mu) \\
&+ \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu}-\frac{(\mu-\mu')}{\sigma}\right)^2\right] & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 \\ \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2 \end{cases}
\end{aligned}
$$

$$
F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi\left(\sqrt{\tilde{q}_\mu}-\frac{(\mu-\mu')}{\sigma}\right) & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2\ , \\ \Phi\left(\frac{\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2}{2\mu/\sigma}\right) & \tilde{q}_\mu > \mu^2/\sigma^2\ . \end{cases}
$$

Monte Carlo test of asymptotic formulae

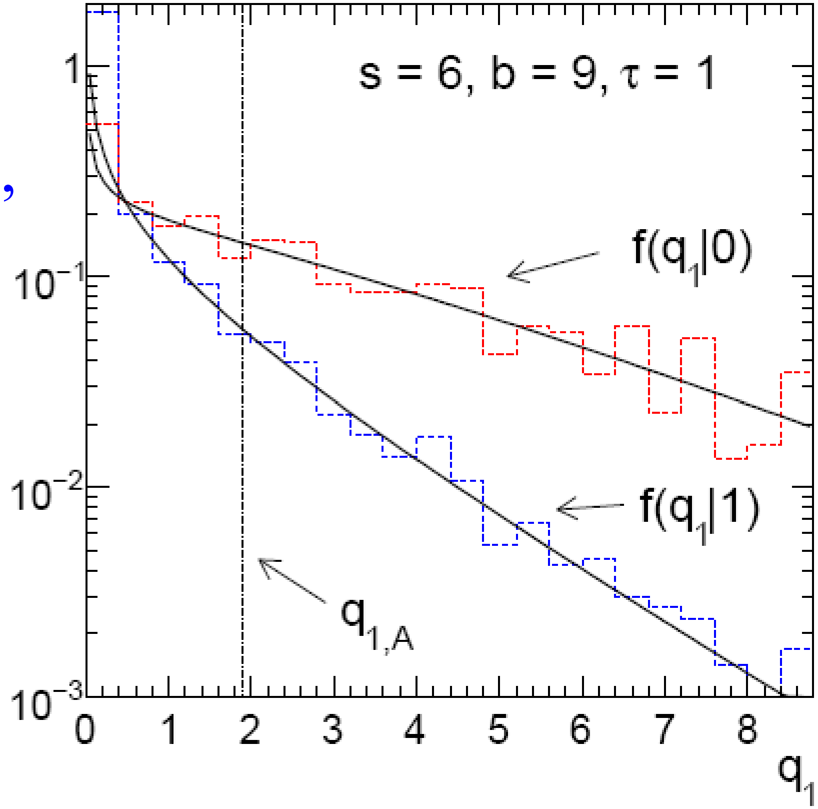
Consider again $n \sim$ Poisson $(\mu s + b)$, $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median$[q_1|0]$ gives "exclusion sensitivity".

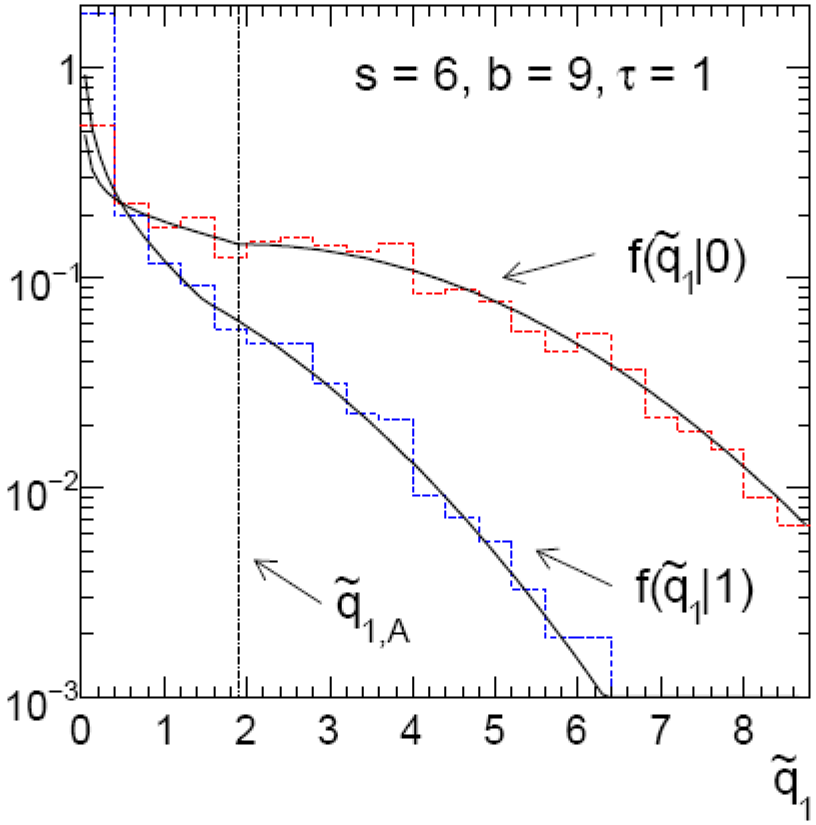
Here asymptotic formulae good for $s = 6$, $b = 9$.



Monte Carlo test of asymptotic formulae

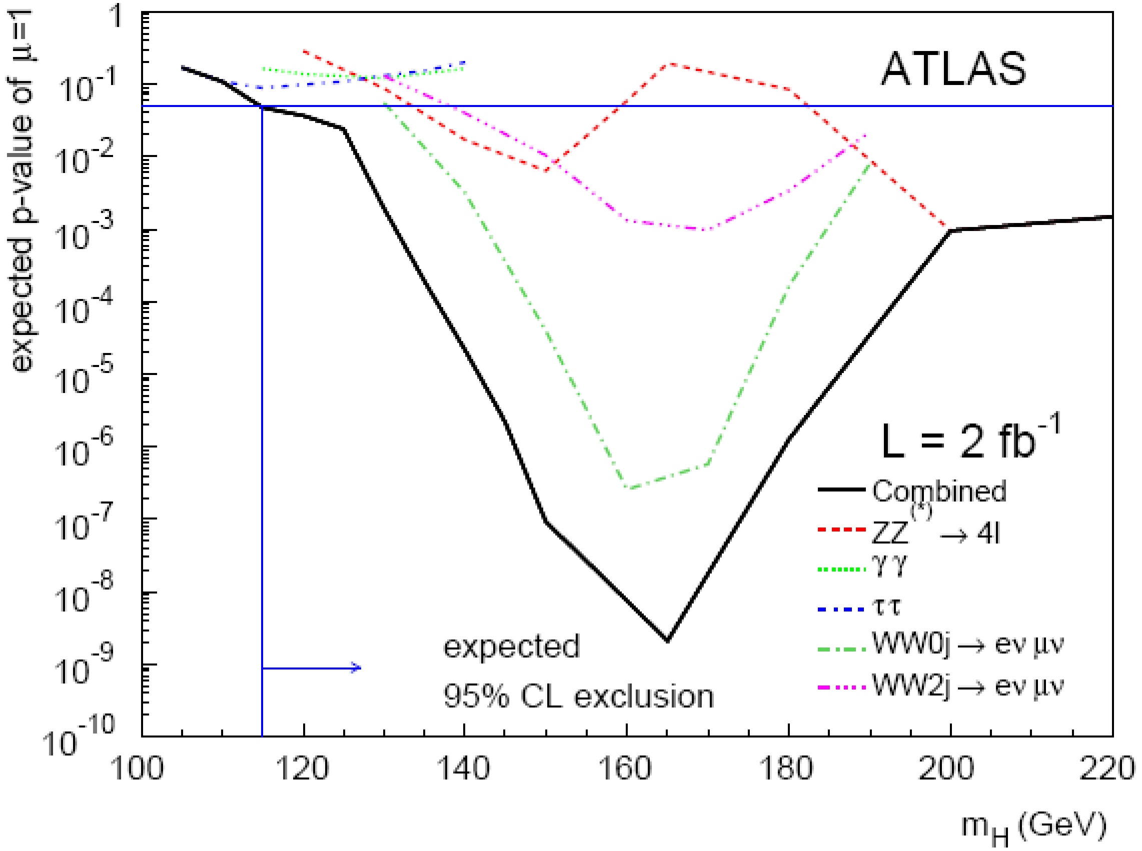
Same message for test based on $\tilde{q}_\mu$.

q_μ and $\tilde{q}_\mu$ give similar tests to the extent that asymptotic formulae are valid.



Example: exclusion sensitivity

Median *p*-value of $\mu = 1$ hypothesis versus Higgs mass assuming background-only data (ATLAS, arXiv:0901.0512).



Summary on discovery and limits

Test values of the parameter(s) of interest; result is *p*-value

If $p < \alpha$, reject hypothesized parameter

Rejecting the background-only hypothesis ≈ discovery

Confidence interval for parameter at CL = 1 – α is range of values not rejected in test of size α.

Test can be based on likelihood ratio (or some approximation)

Systematic uncertainties ↔ nuisance parameters

“Tevatron Style”: $Q = -2\ln(L_{s+b}/L_b)$

Profile Likelihood Ratio: $q_\mu = -2\ln(L(\mu,\hat{\hat{\theta}})/L(\hat{\mu},\hat{\theta}))$

Can (should) also use Bayesian methods (no time for this today)

Extra slides

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
$P\,(0.117 < \alpha_s < 0.121)$,

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

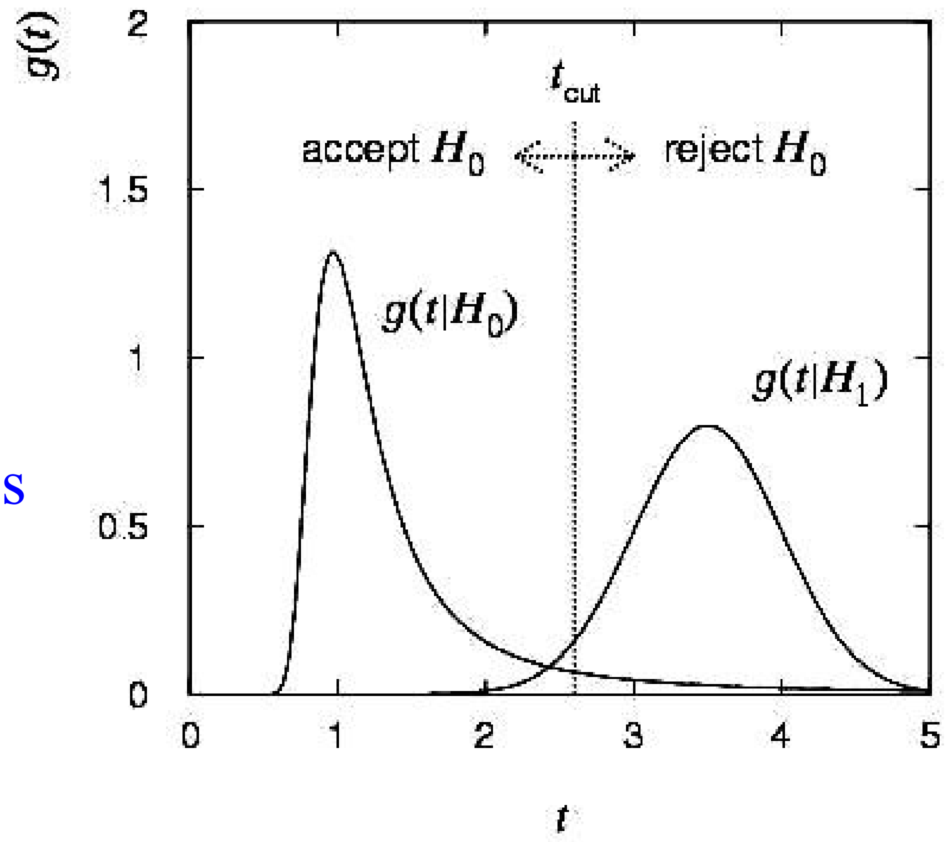
No golden rule for priors (“if-then” character of Bayes’ thm.)

Significance level and power

Probability to reject H_0 if it is true (type-I error):

$$\alpha = \int_{t_{\rm cut}}^{\infty} g(t|H_0)\, dt$$

(significance level)



Probability to accept H_0 if H_1 is true (type-II error):

$$\beta = \int_{-\infty}^{t_{\rm cut}} g(t|H_1)\, dt$$

($1 - \beta =$ power)

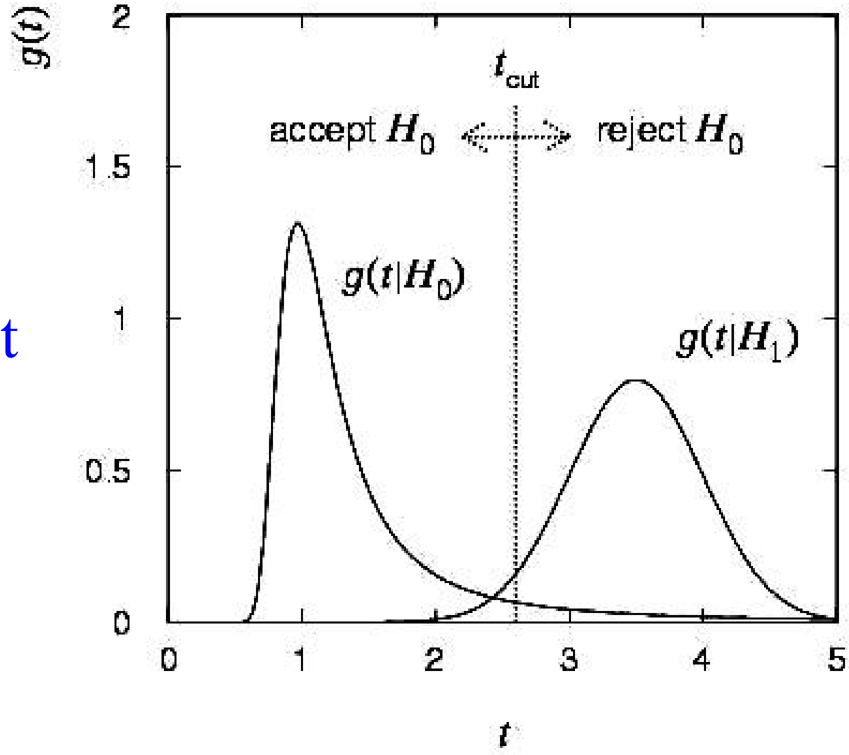
Signal/background efficiency

Probability to reject background hypothesis for background event (background efficiency):

$$\varepsilon_{\rm b} = \int_{t_{\rm cut}}^{\infty} g(t|{\rm b})\, dt = \alpha$$

Probability to accept a signal event as signal (signal efficiency):

$$\varepsilon_{\rm s} = \int_{t_{\rm cut}}^{\infty} g(t|{\rm s})\, dt = 1 - \beta$$



Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are π_s and π_b (prior probabilities).

Suppose we select signal events with $t > t_{cut}$. What is the 'purity' of our selected sample?

Here purity means the probability to be signal given that the event was accepted. Using Bayes' theorem we find:

$$P(\mathrm{s}|t > t_{\mathrm{cut}}) = \frac{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}}}{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}} + P(t > t_{\mathrm{cut}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= \frac{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}}}{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}} + \varepsilon_{\mathrm{b}}\pi_{\mathrm{b}}}$$

So the purity depends on the prior probabilities as well as on the signal and background efficiencies.

Wald approximation for profile likelihood ratio

To find *p*-values, we need: $f(q_0|0), \quad f(q_\mu|\mu)$

For median significance under alternative, need: $f(q_\mu|\mu')$

Use approximation due to Wald (1943)

$$-2\ln\lambda(\mu) = \frac{(\mu - \hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N})$$

$$\hat{\mu} \sim \text{ Gaussian}(\mu', \sigma)$$

sample size

$$\text{i.e., } E[\hat{\mu}] = \mu'$$

σ from covariance matrix V, use, e.g.,

$$V^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

Noncentral chi-square for $-2\ln\lambda(\mu)$

If we can neglect the $O(1/\sqrt{N})$ term, $-2\ln\lambda(\mu)$ follows a noncentral chi-square distribution for one degree of freedom with noncentrality parameter

$$\Lambda = \frac{(\mu - \mu')^2}{\sigma^2}$$

As a special case, if $\mu' = \mu$ then $\Lambda = 0$ and $-2\ln\lambda(\mu)$ follows a chi-square distribution for one degree of freedom (Wilks).

The Asimov data set

To estimate median value of $-2\ln\lambda(\mu)$, consider special data set where all statistical fluctuations suppressed and n_i, m_i are replaced by their expectation values (the “Asimov” data set):

$$\begin{aligned} n_i &= \mu' s_i + b_i \\ m_i &= u_i \end{aligned}$$

$$\longrightarrow \quad \hat{\mu} = \mu' \qquad \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$$

$$\lambda_{\mathrm{A}}(\mu) = \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\mathrm{A}}(\hat{\mu}, \hat{\boldsymbol{\theta}})} = \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_A(\mu', \boldsymbol{\theta})}$$

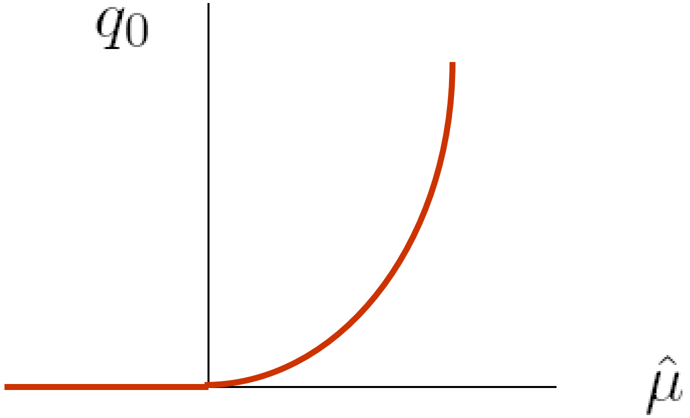
$$-2\ln\lambda_{\mathrm{A}}(\mu) = \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

Asimov value of $-2\ln\lambda(\mu)$ gives non-centrality param. Λ, or equivalently, σ

Relation between test statistics and $\hat{\mu}$

Assuming Wald approximation, the relation between q_0 and $\hat{\mu}$ is

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$



Monotonic, therefore quantiles of $\hat{\mu}$ map one-to-one onto those of q_0, e.g.,

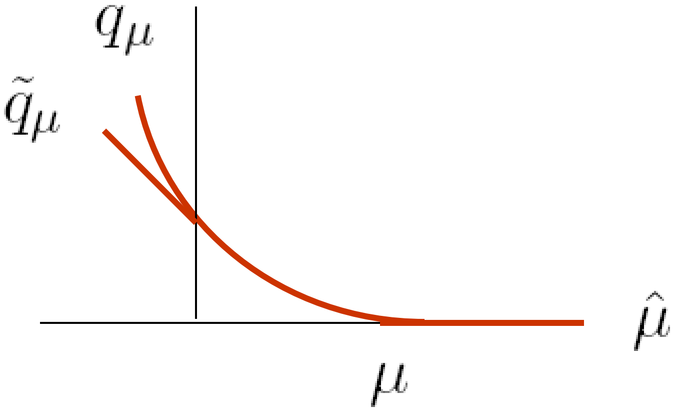
$$\mathrm{med}[q_0] = q_0(\mathrm{med}[\hat{\mu}]) = q_0(\mu') = \frac{\mu'^2}{\sigma^2} = -2\ln\lambda_{\mathrm{A}}(0)$$

$$\mathrm{med}[Z_0] = \sqrt{-2\ln\lambda_{\mathrm{A}}(0)}$$

Relation between test statistics and $\hat{\mu}$

Assuming the Wald approximation for $-2\ln\lambda(\mu)$, q_μ and $\tilde{q}_\mu$ both have monotonic relation with μ.

$$q_\mu = \begin{cases} \frac{(\mu-\hat{\mu})^2}{\sigma^2} & \hat{\mu} < \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$



$$\tilde{q}_\mu = \begin{cases} \frac{\mu^2}{\sigma^2} - \frac{2\mu\hat{\mu}}{\sigma^2} & \hat{\mu} < 0 \\ \frac{(\mu-\hat{\mu})^2}{\sigma^2} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \ , \end{cases}$$

And therefore quantiles of q_μ, $\tilde{q}_\mu$ can be obtained directly from those of $\hat{\mu}$ (which is Gaussian).

Combination of channels

For a set of independent decay channels, full likelihood function is product of the individual ones:

$$L(\mu, \boldsymbol{\theta}) = \prod_i L_i(\mu, \boldsymbol{\theta}_i)$$

For combination need to form the full function and maximize to find estimators of μ, $\boldsymbol{\theta}$.

→ ongoing ATLAS/CMS effort with RooStats framework

Trick for median significance: estimator for μ is equal to the Asimov value μ' for all channels separately, so for combination,

$$\lambda_{\mathrm{A}}(\mu) = \prod_i \lambda_{\mathrm{A},i}(\mu) \qquad \text{where} \qquad \lambda_{\mathrm{A},i}(\mu) = \frac{L_i(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_i(\mu', \boldsymbol{\theta})}$$

Discovery significance for $n \sim \text{Poisson}(s + b)$

Consider again the case where we observe n events, model as following Poisson distribution with mean $s + b$ (assume b is known).

1) For an observed n, what is the significance Z_0 with which we would reject the $s = 0$ hypothesis?

2) What is the expected (or more precisely, median) Z_0 if the true value of the signal rate is s?

Gaussian approximation for Poisson significance

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

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

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

Significance for rejecting $s = 0$ is therefore

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

Expected (median) significance assuming signal rate s is

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

Better approximation for Poisson significance

Likelihood function for parameter *s* is

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

or equivalently the log-likelihood is

$$\ln L(s) = n \ln(s+b) - (s+b) - \ln n!$$

Find the maximum by setting $\frac{\partial \ln L}{\partial s} = 0$

gives the estimator for *s*: $\hat{s} = n - b$

Approximate Poisson significance (continued)

The likelihood ratio statistic for testing $s = 0$ is

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

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

$$Z_0 \approx \sqrt{q_0} = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \quad \text{for } n > b, \ 0 \text{ otherwise}$$

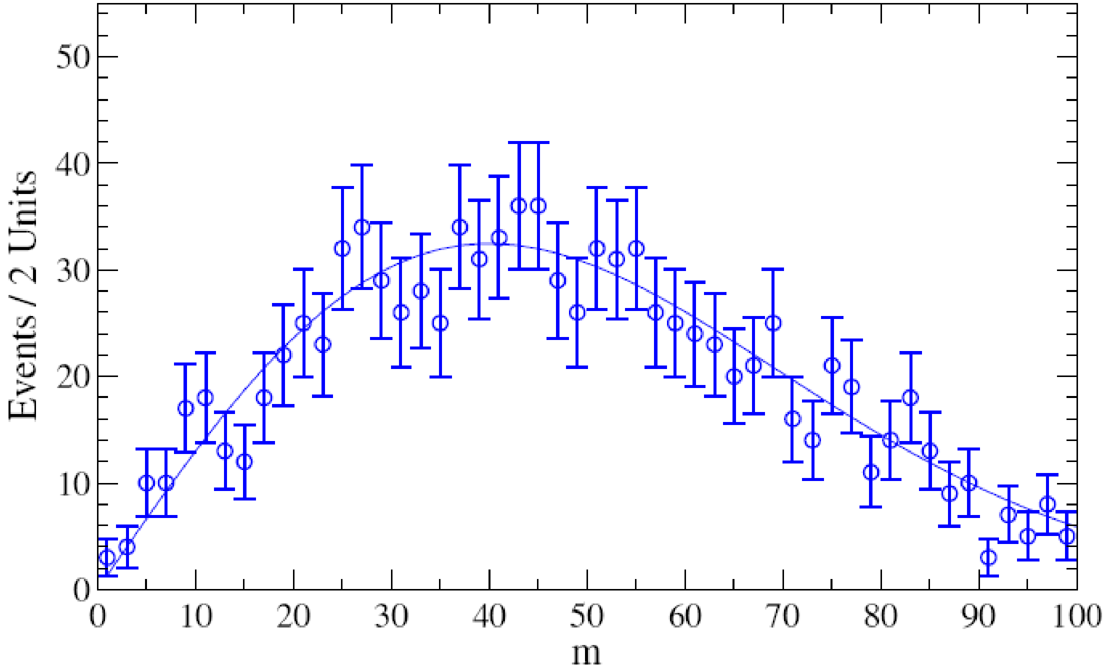
To find median[$Z_0|s+b$], let $n \rightarrow s + b$,

$$\boxed{\text{median}[Z_0|s+b] \approx \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}}$$

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

Example 2: Shape analysis

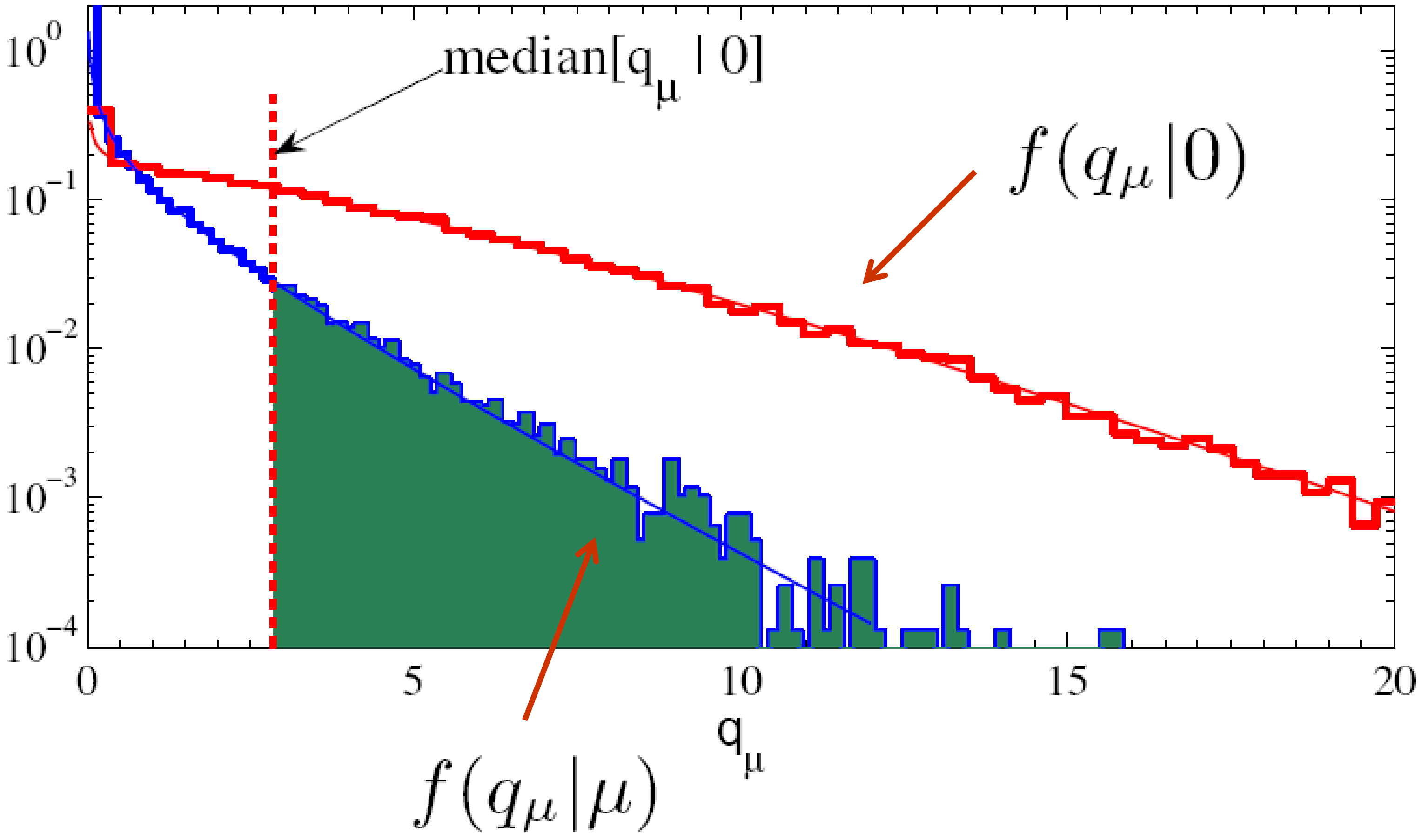
Look for a Gaussian bump sitting on top of:



$$L(\mu, \theta) = \prod_{i=1}^{N} \frac{(\mu s_i + \theta f_{\mathrm{b},i})^{n_i}}{n_i!} e^{-(\mu s_i + \theta f_{\mathrm{b},i})}$$

Monte Carlo test of asymptotic formulae

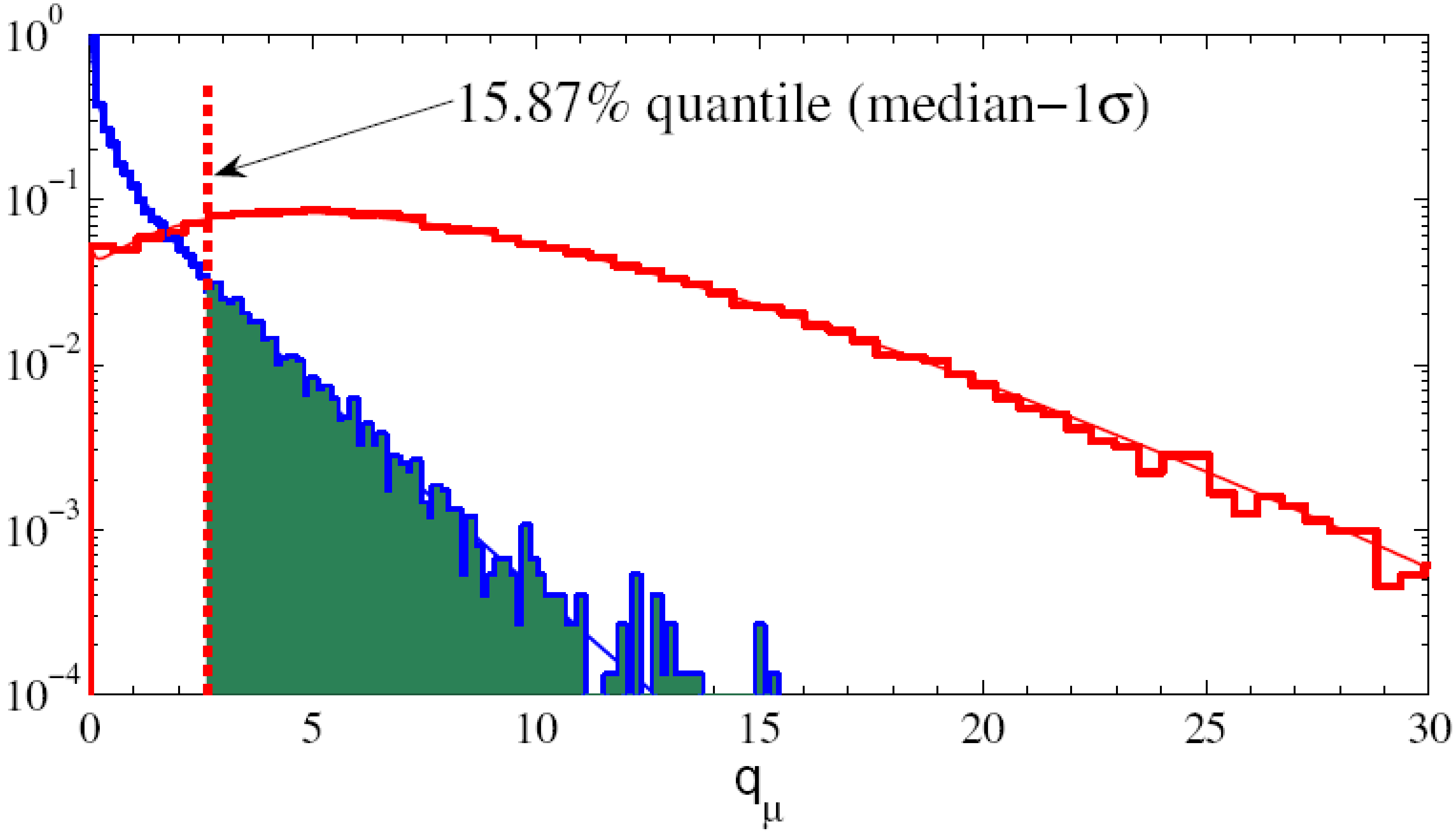
Distributions of q_μ here for μ that gave $p_\mu = 0.05$.



Using $f(q_\mu|0)$ to get error bands

We are not only interested in the median[qm|0]; we want to know how much statistical variation to expect from a real data set.

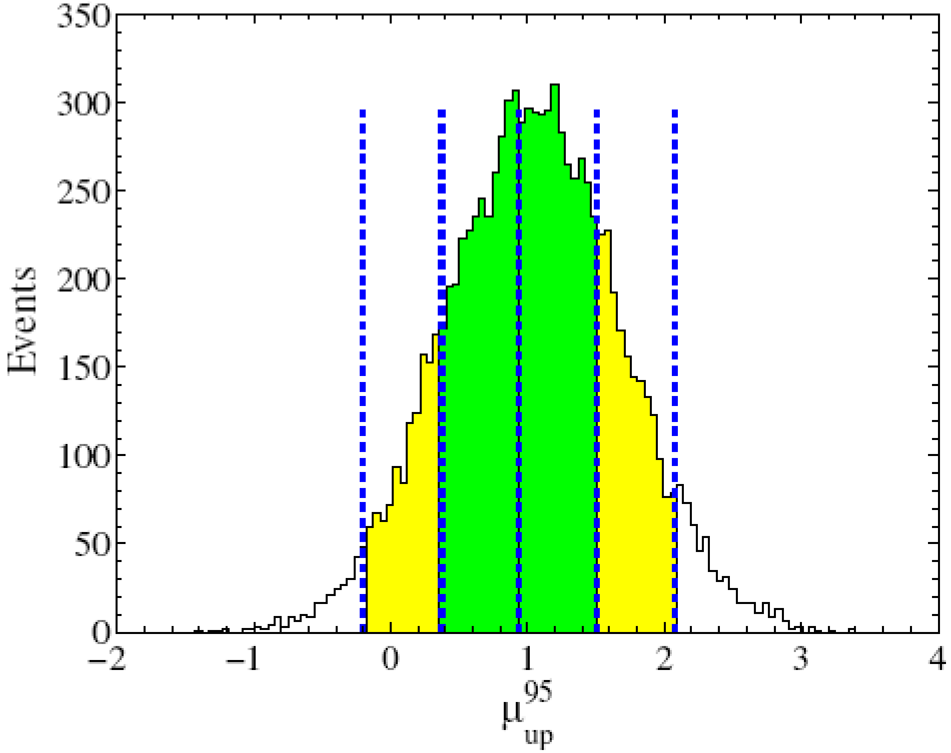
But we have full $f(q_\mu|0)$; we can get any desired quantiles.



Distribution of upper limit on μ

$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from MC;

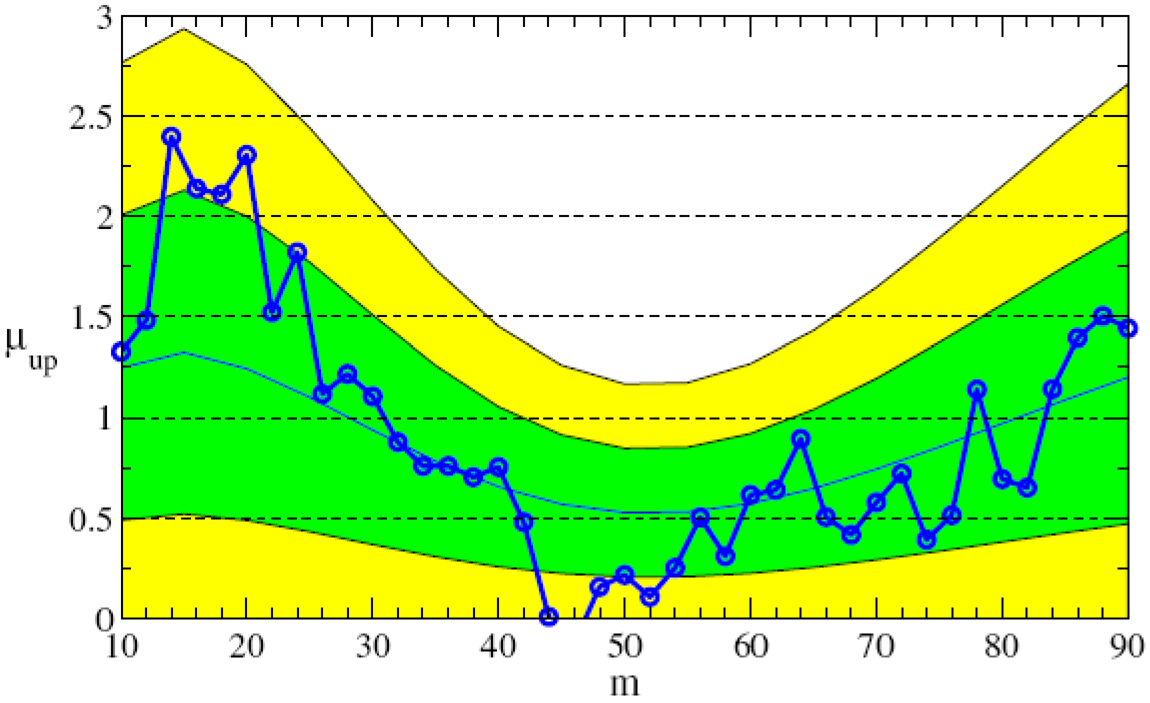
Vertical lines from asymptotic formulae



Limit on μ versus peak position (mass)

$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from asymptotic formulae;

Points are from a single arbitrary data set.


μ_{up}
m