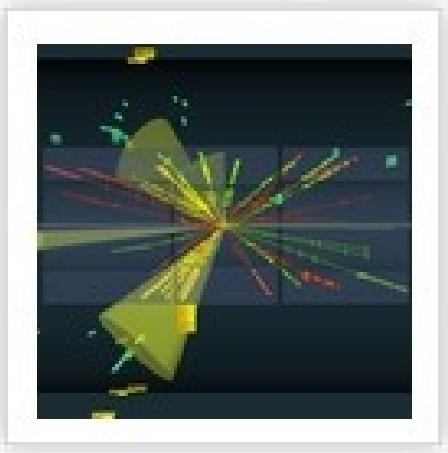
Statistical Methods for HEP

Lecture 3: Discovery and Limits



Taller de Altas Energías 2010

Universitat de Barcelona

1-11 September 2010

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

Outline

Lecture 1: Introduction and basic formalism
Probability, statistical tests, parameter estimation.

Lecture 2: Multivariate methods
General considerations
Example of a classifier: Boosted Decision Trees

→ Lecture 3: Discovery and Exclusion limits
Quantifying significance and sensitivity
Systematic uncertainties (nuisance parameters)

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.

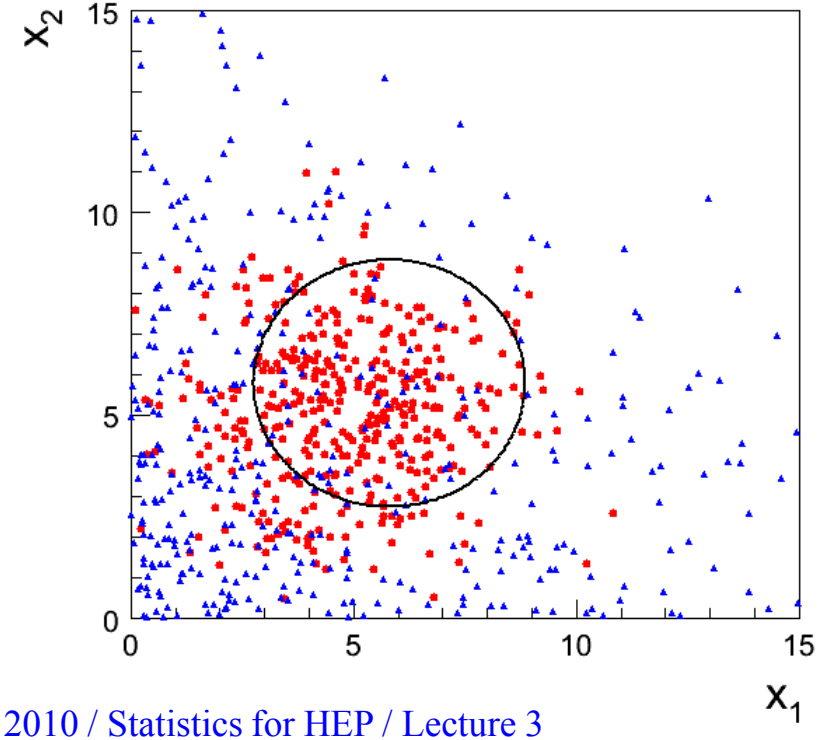
Likelihood ratio as test statistic

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.

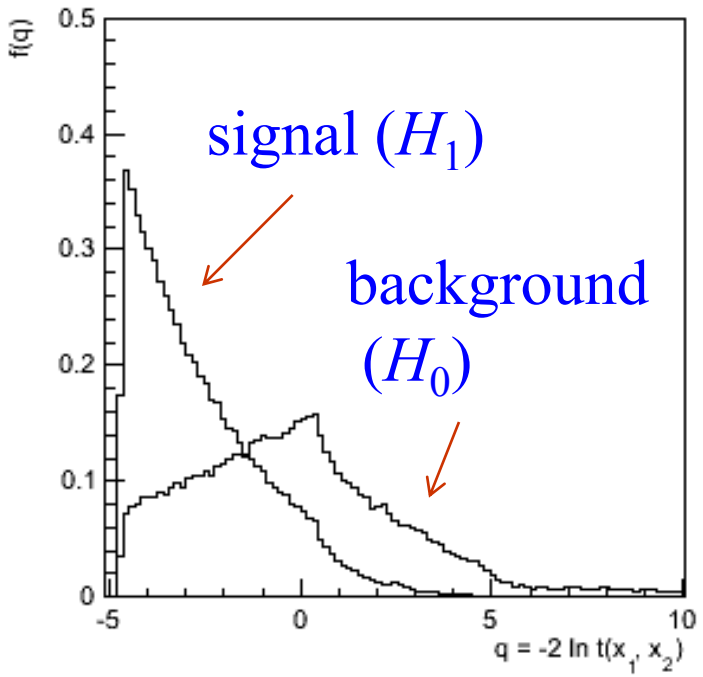
But in this example we can find contours of constant likelihood ratio such as:



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:

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

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

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

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

Note that the term $-s$ in front is a constant and can be dropped.

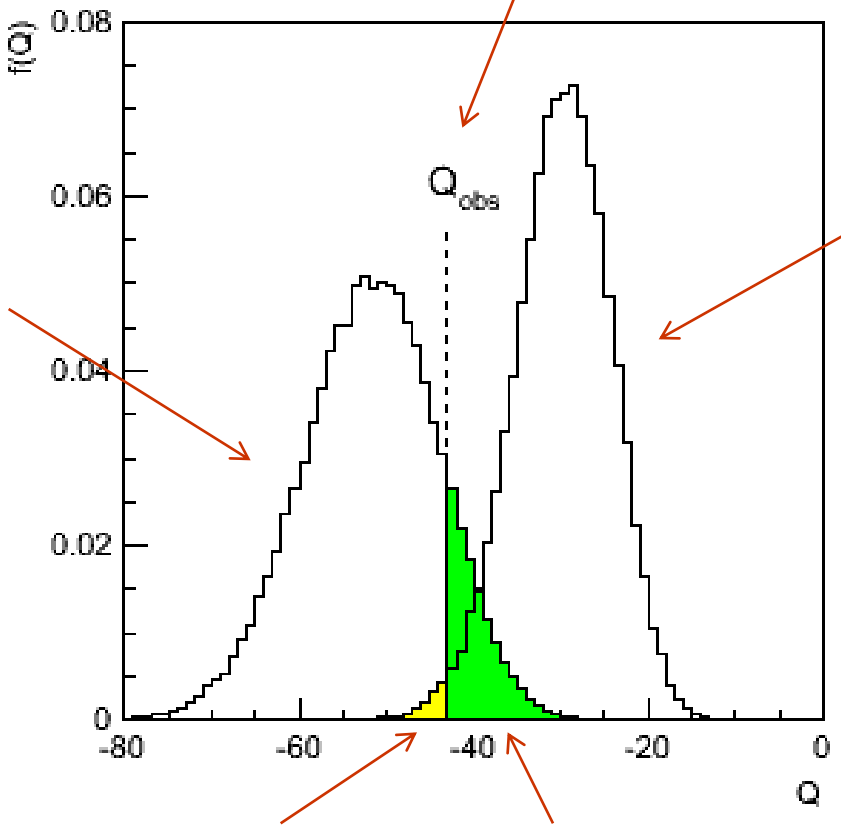
The rest is a sum of contributions for each event, and each term in the sum has the same distribution.

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.



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

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

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:

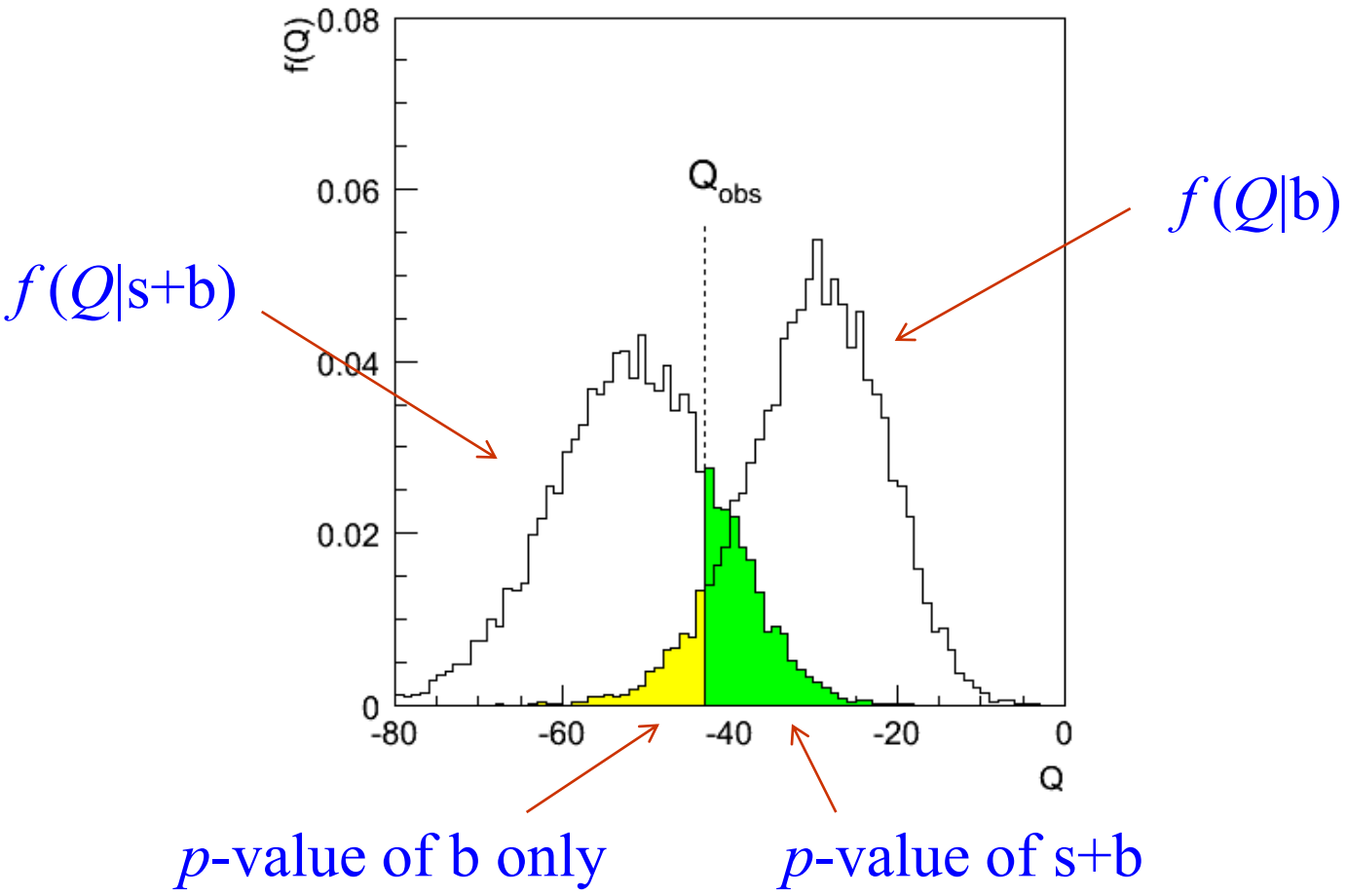
$$f(Q) = \int f(Q|b)\pi(b)\,db$$

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



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

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

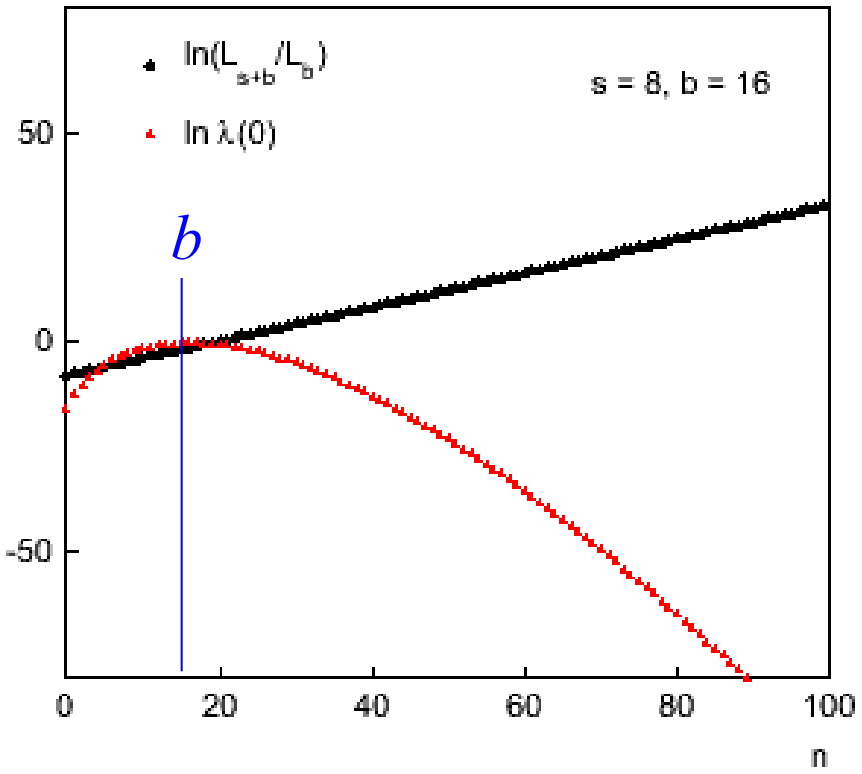
whereas previously we had used

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

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).

Prototype search analysis

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

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

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

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 ($\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$,b_{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:

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

maximizes L for specified μ

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

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

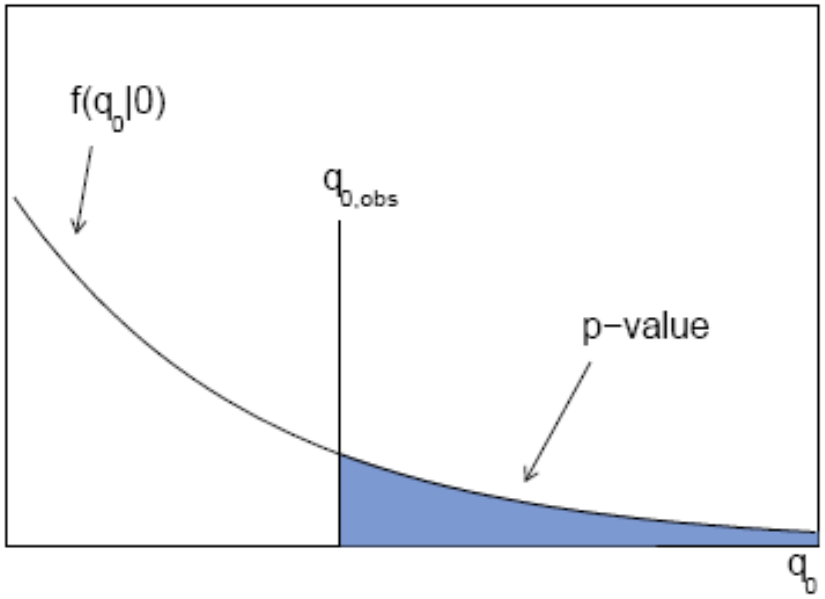
i.e. only regard upward fluctuation of data as evidence against the background-only hypothesis.

p-value for discovery

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

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

will get formula for this later

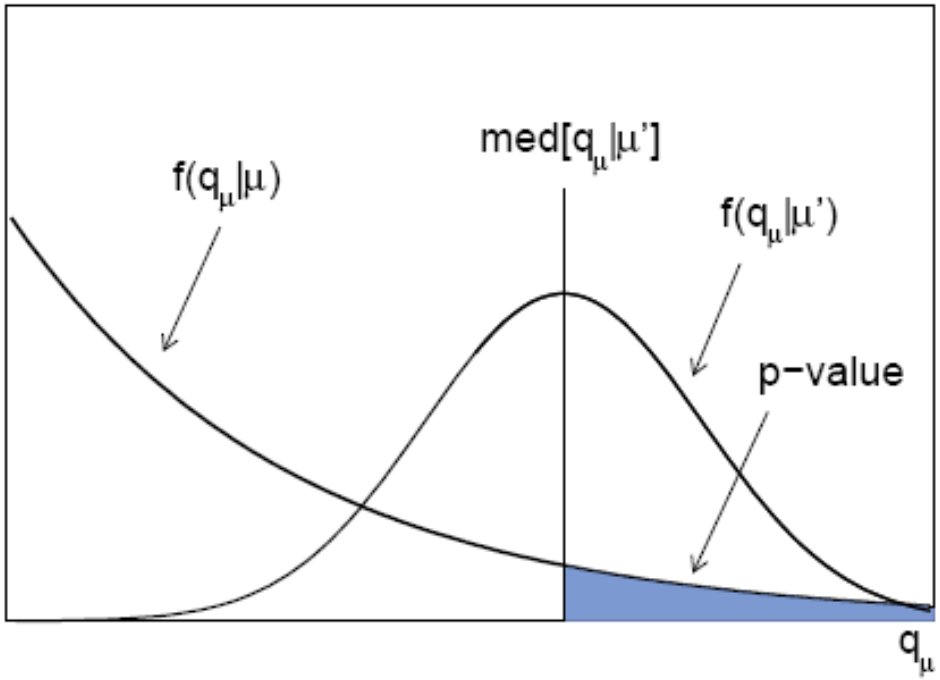


From *p*-value get equivalent significance,

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

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 $f(q_0|\mu')$,

Wald approximation for profile likelihood ratio

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells, *Using the Profile Likelihood in Searches for New Physics*, in preparation.

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

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).

Distribution of q_0

Assuming the Wald approximation, we can write down the full distribution of q_0 as

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

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

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

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

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

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):

$$n_i = \mu' s_i + b_i$$
$$m_i = u_i$$

$\rightarrow$ $\hat{\mu}_{\rm A} = \mu'$, $\boldsymbol{\theta}_{\rm A} = \hat{\boldsymbol{\theta}}$ from very large MC sample, so

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

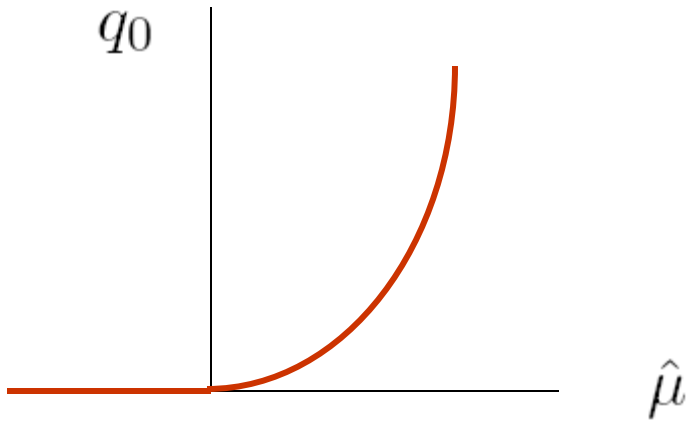
$$-2\ln\lambda_{\rm A}(\mu) = \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda$$

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.,

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

$$\text{med}[Z_0] = \sqrt{-2\ln\lambda_A(0)}$$

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):

$H \to \gamma\gamma$

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

$H \to ZZ^{(*)} \to 4l$ (l = e, μ)

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

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.

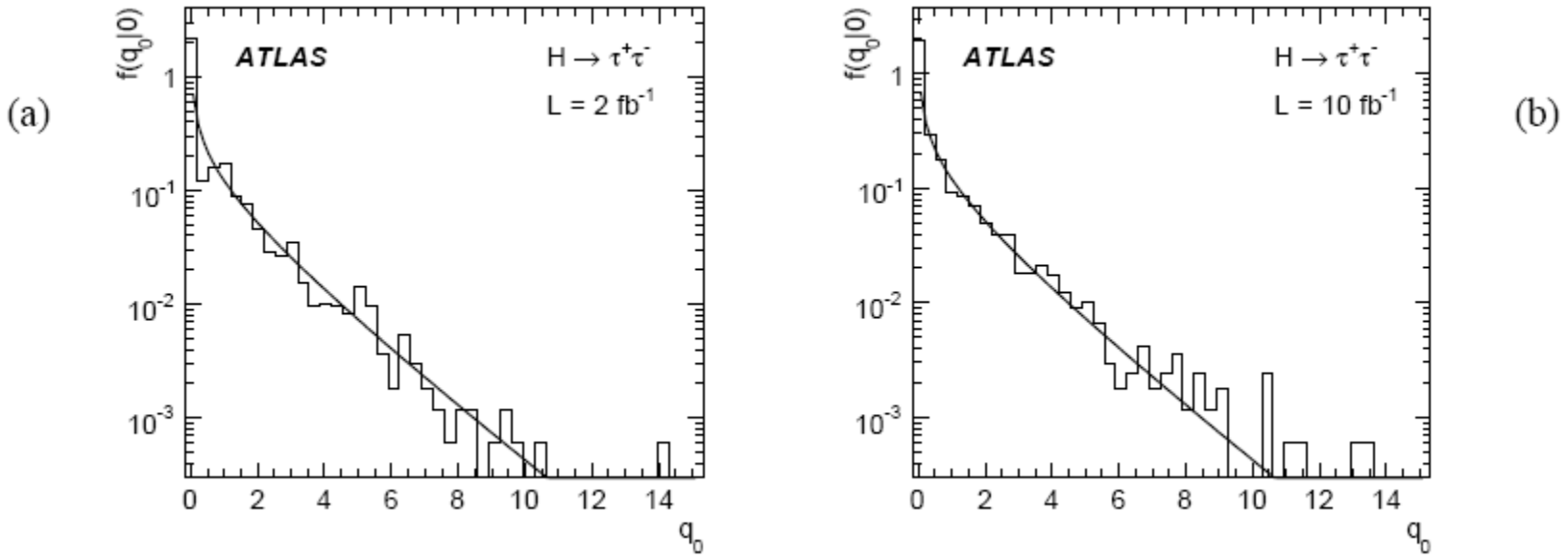


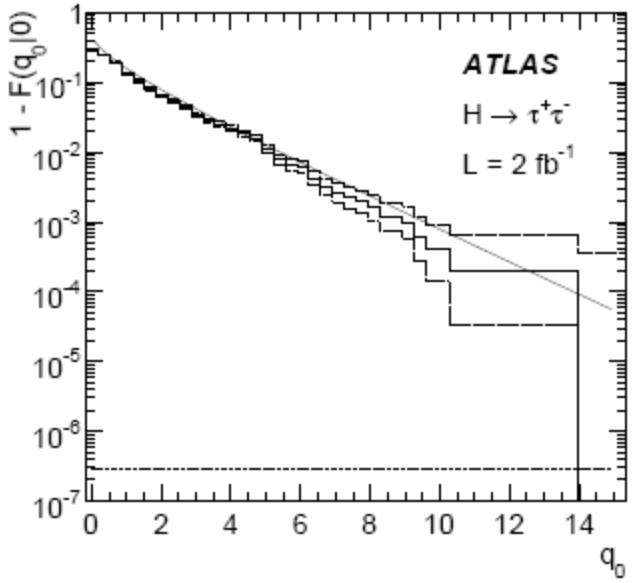
Figure 12: The distribution of the test statistic q_0 for $H \to \tau^+\tau^-$ under the null background-only hypothesis, for $m_H = 130\,\text{GeV}$ with an integrated luminosity of 2 (a) and 10 (b) fb^{-1}. A $\frac{1}{2}\chi^2_1$ distribution is superimposed. Figures (c) and (d) show $1 - F(q_0)$ 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.

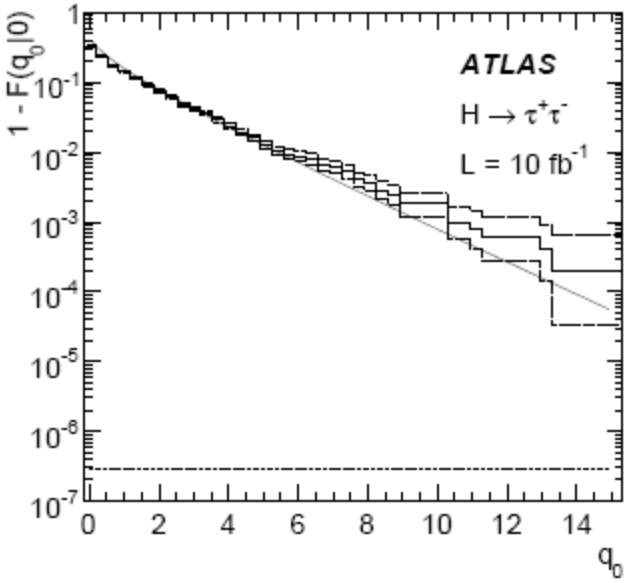
Cumulative distributions of q_0

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.

(c)





(d)

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, $\boldsymbol{\theta}$.

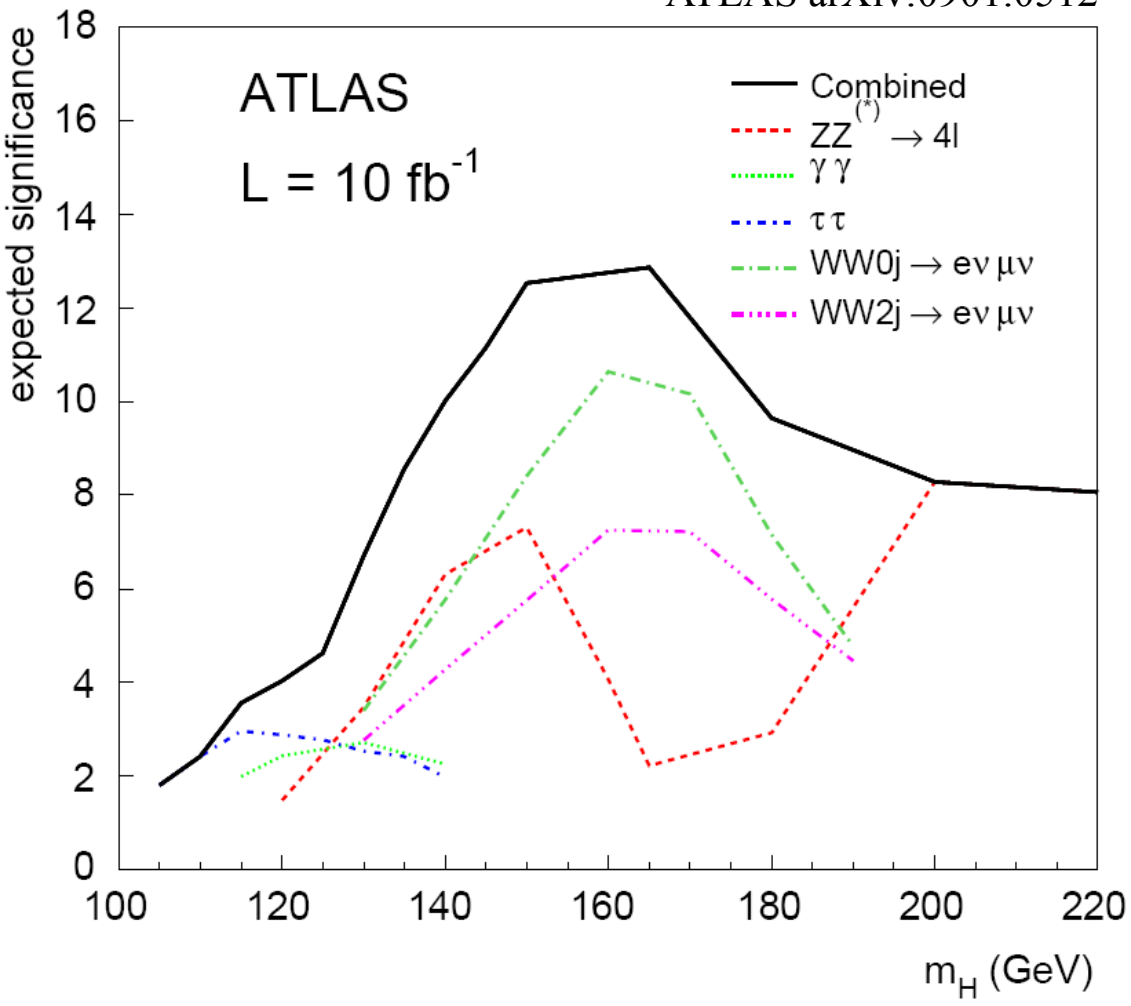
→ ongoing ATLAS/CMS effort with RooStats framework

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

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

Combined median significance

ATLAS arXiv:0901.0512



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

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

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

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

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

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 $\dfrac{\partial \ln L}{\partial s} = 0$

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 $n \to s + b$,

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

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

Wrapping up lecture 2

To establish discovery, we need (at least) to reject the background-only hypothesis (usually with $Z > 5$).

For two simple hypotheses, best test of H_0 is from likelihood ratio $L(H_1)/L(H_0)$.

Alternative likelihood ratio $\lambda(\mu) = L(\mu) / L(\hat{\mu})$ will lead to similar test; can use Wilks' theorem to obtain sampling distribution.

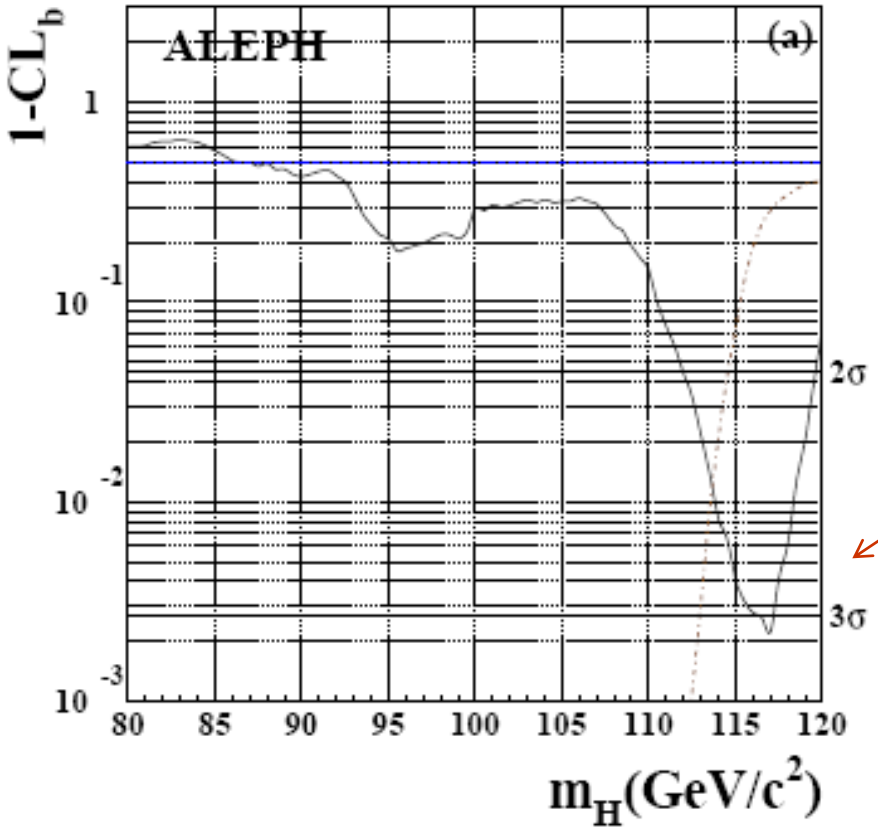
Methods for incorporating systematics somewhat different for the two approaches (in practice lead to similar results).

Next: exclusion limits

Extra slides

Example: ALEPH Higgs search

p-value $(1 - CL_b)$ of background only hypothesis versus tested Higgs mass measured by ALEPH Experiment

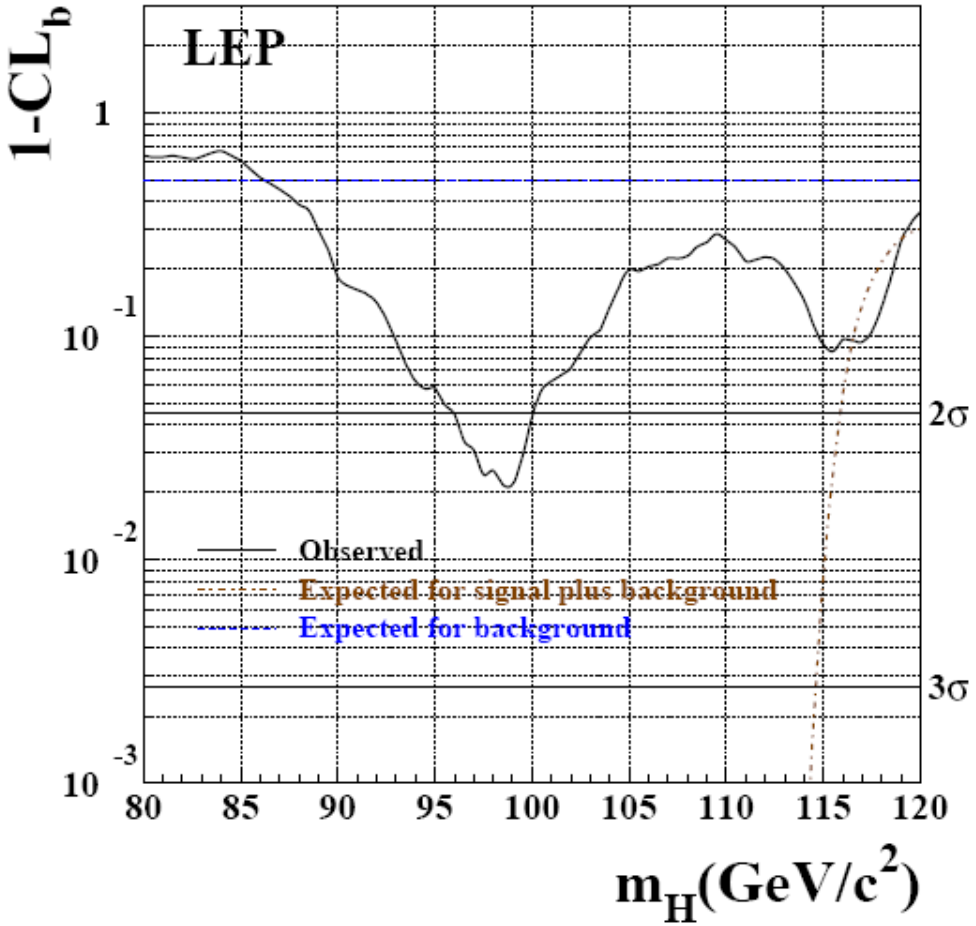


Possible signal?

Phys.Lett.B565:61-75,2003.
hep-ex/0306033

Example: LEP Higgs search

Not seen by the other LEP experiments. Combined analysis gives p-value of background-only hypothesis of 0.09 for $m_H = 115$ GeV.



Phys.Lett.B565:61-75,2003.
hep-ex/0306033