Some Statistical Tools for Particle Physics

Particle Physics Colloquium
MPI für Physik u. Astrophysik
Munich, 10 May, 2016

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

Outline

1) Brief review of HEP context and statistical tests.

2) Statistical tests based on the profile likelihood ratio

3) A measure of discovery sensitivity is often used to plan a future analysis, e.g., $s/\sqrt{b}$, gives approximate expected discovery significance (test of $s = 0$) when counting $n \sim \text{Poisson}(s+b)$. A measure of discovery significance is proposed that takes into account uncertainty in the background rate.

4) Brief comment on importing tools from Machine Learning & choice of variables for multivariate analysis

Data analysis in particle physics

Particle physics experiments are expensive

e.g. LHC, ~ $\$10^{10}$ (accelerator and experiments)

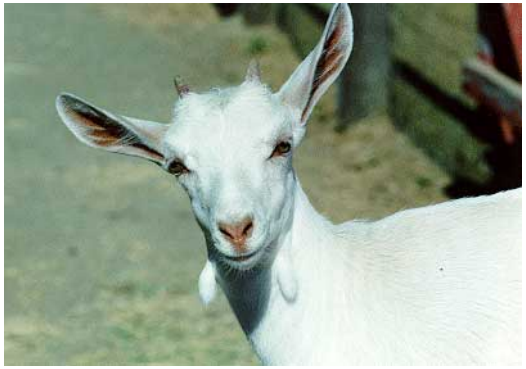
the competition is intense

(ATLAS vs. CMS) vs. many others

and the stakes are high:

4 sigma effect

5 sigma effect



Hence the increased interest in advanced statistical methods.

Prototypical HEP 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",...)

(Frequentist) statistical tests

Consider test of a parameter μ, e.g., proportional to cross section.

Result of measurement is a set of numbers $\boldsymbol{x}$.

To define test of μ, specify *critical region* w_μ, such that probability to find $\boldsymbol{x} \in w_\mu$ is not greater than α (the *size* or *significance level*):

$$P(\mathbf{x} \in w_\mu | \mu) \leq \alpha$$

(Must use inequality since $\boldsymbol{x}$ may be discrete, so there may not exist a subset of the data space with probability of exactly α.)

Equivalently define a p-value p_μ equal to the probability, assuming μ, to find data at least as "extreme" as the data observed.

The critical region of a test of size α can be defined from the set of data outcomes with $p_\mu < \alpha$. Often use, e.g., $\alpha = 0.05$.

If observe $\boldsymbol{x} \in w_\mu$, reject μ.

Test statistics and p-values

Often construct a scalar test statistic, $q_\mu(\boldsymbol{x})$, which reflects the level of agreement between the data and the hypothesized value μ.

For examples of statistics based on the profile likelihood ratio, see, e.g., CCGV, EPJC 71 (2011) 1554; arXiv:1007.1727.

Usually define q_μ such that higher values represent increasing incompatibility with the data, so that the p-value of μ is:

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

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu < \alpha$.

Confidence interval from inversion of a test

Carry out a test of size α for all values of μ.

The values that are not rejected constitute a *confidence interval* for μ at confidence level CL = $1 - \alpha$.

> The confidence interval will by construction contain the true value of μ with probability of at least $1 - \alpha$.

The interval depends on the choice of the critical region of the test.

Put critical region where data are likely to be under assumption of the relevant alternative to the μ that's being tested.

Test $\mu = 0$, alternative is $\mu > 0$: test for discovery.

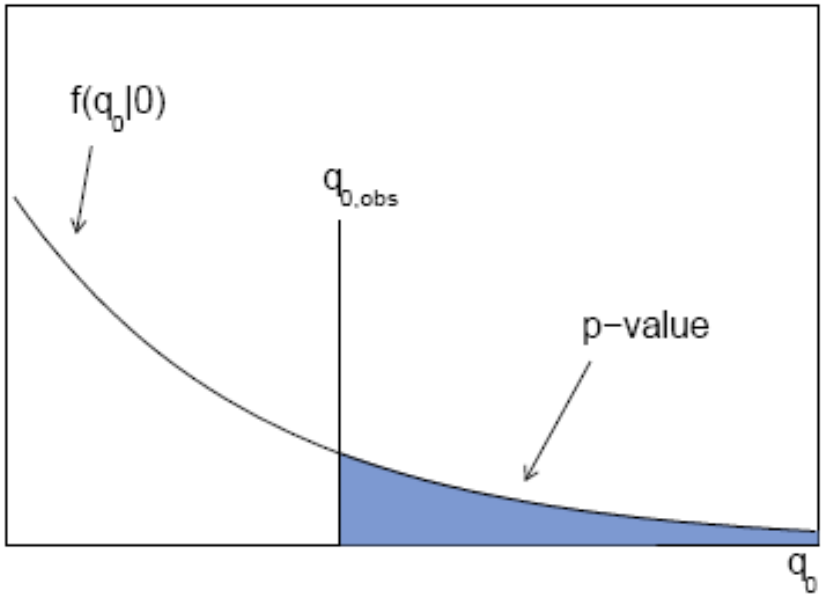
Test $\mu = \mu_0$, alternative is $\mu = 0$: testing all μ_0 gives upper limit.

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

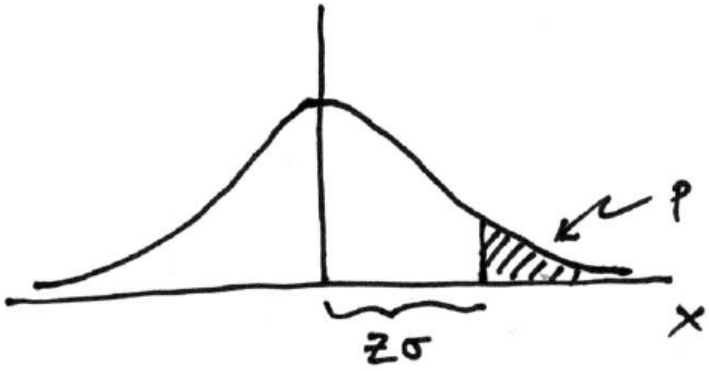


From *p*-value get equivalent significance,

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

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`

Prototype 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

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

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:

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, e.g., $\lambda = L(\mu, \boldsymbol{\theta})/L\,(0, \boldsymbol{\theta})$, gives optimum test (Neyman-Pearson lemma). But the distribution of this statistic depends in general on the nuisance parameters $\boldsymbol{\theta}$, and one can only reject μ if it is rejected for all $\boldsymbol{\theta}$.

The advantage of using the profile likelihood ratio is that the asymptotic (large sample) distribution of $-2\ln \lambda(\mu)$ approaches a chi-square form *independent of the 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 though here physically $\mu \geq 0$, 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.

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

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

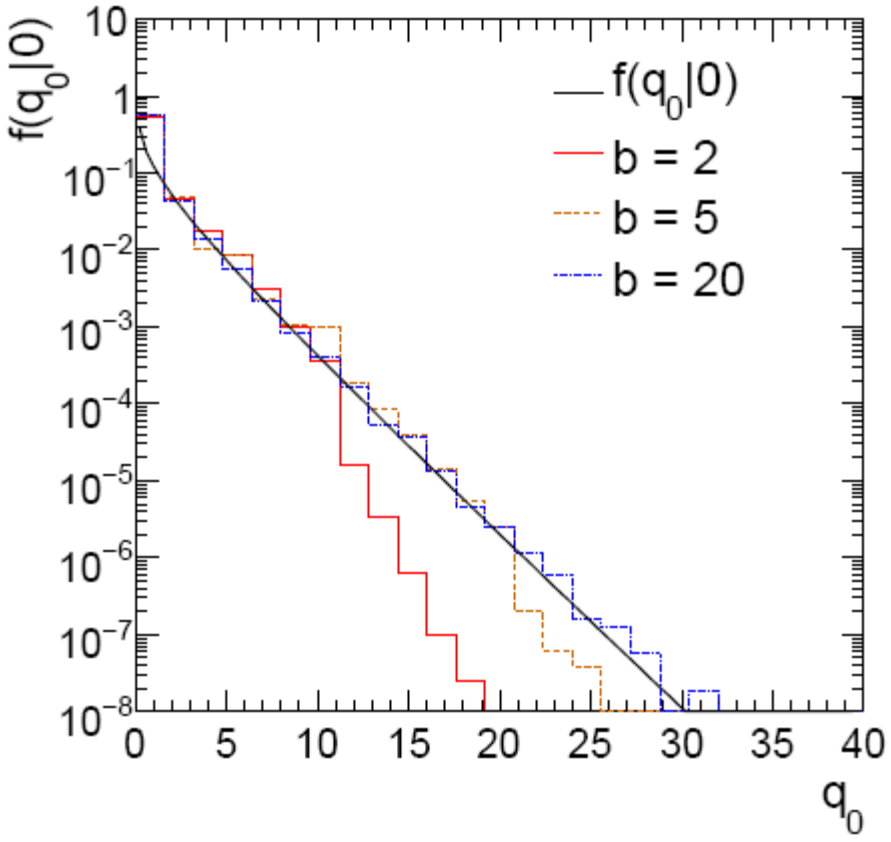
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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

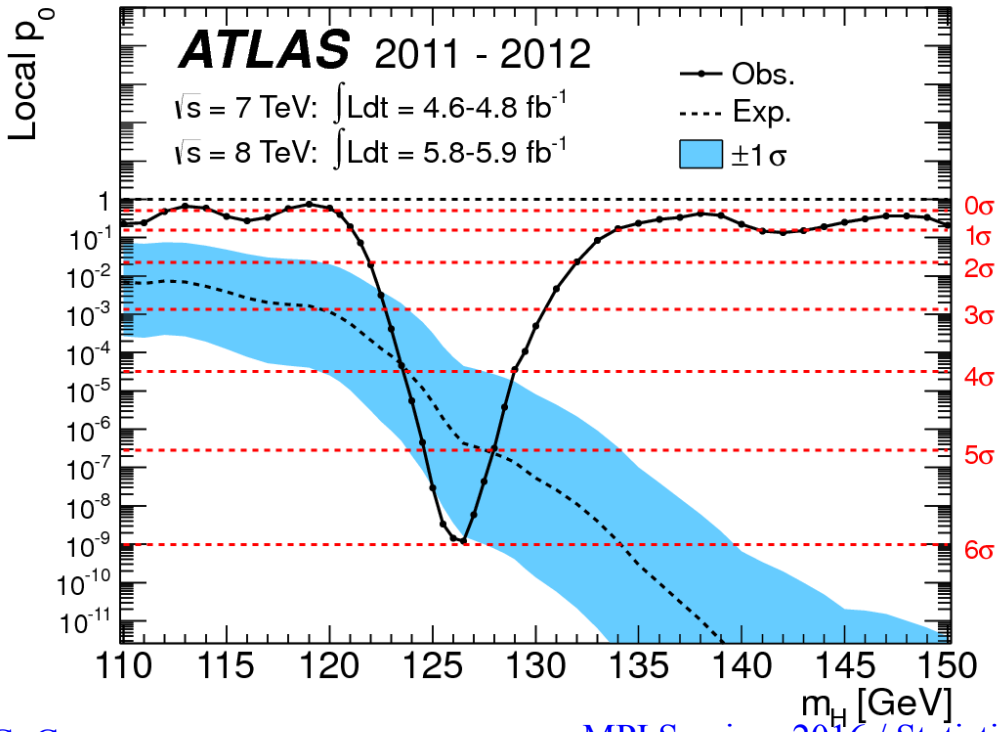


Discovery: the p_0 plot

The "local" p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_H, without any correct for the Look-Elsewhere Effect.

The "Expected" (dashed) curve gives the median p_0 under assumption of the SM Higgs ($\mu = 1$) at each m_H.

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



The blue band gives the width of the distribution ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

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

Independent of nuisance param. in large sample limit

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

Monte Carlo test of asymptotic formulae

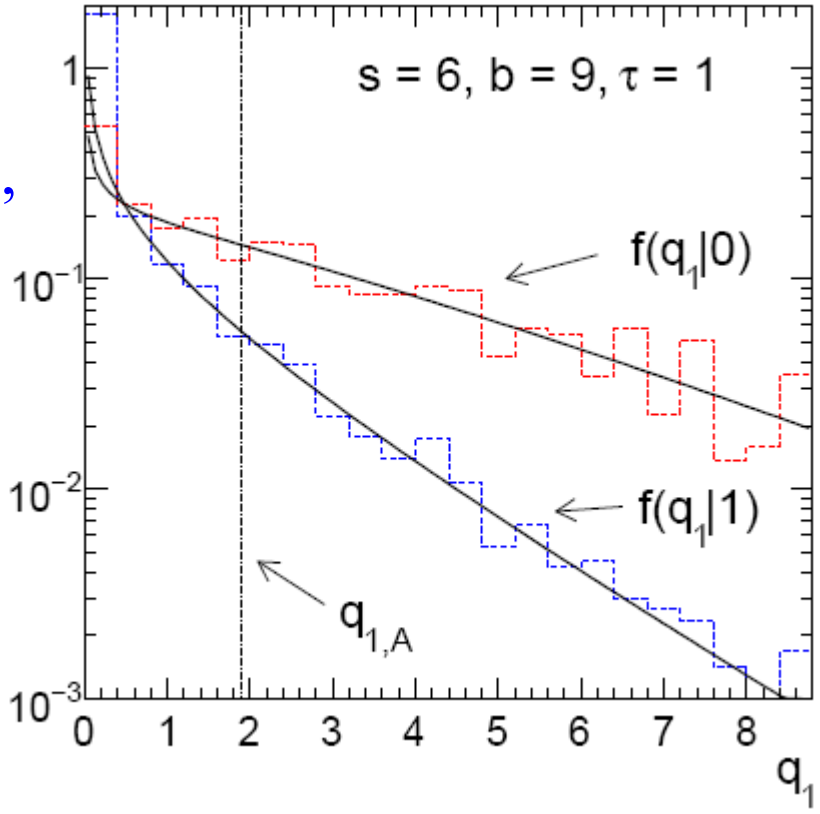
Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of μ=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 $s = 6$, $b = 9$.



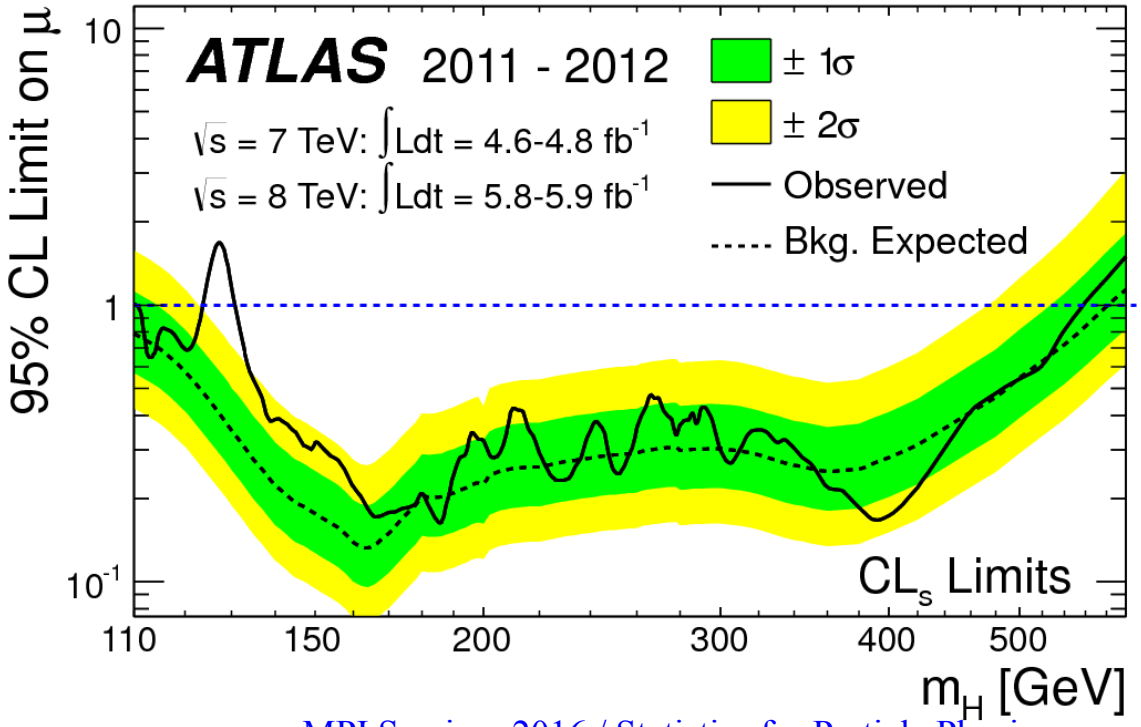
Limits: the "Brazil plot"

For every value of m_H, find the upper limit on μ.

Also for each m_H, determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of $\mu = 0$.

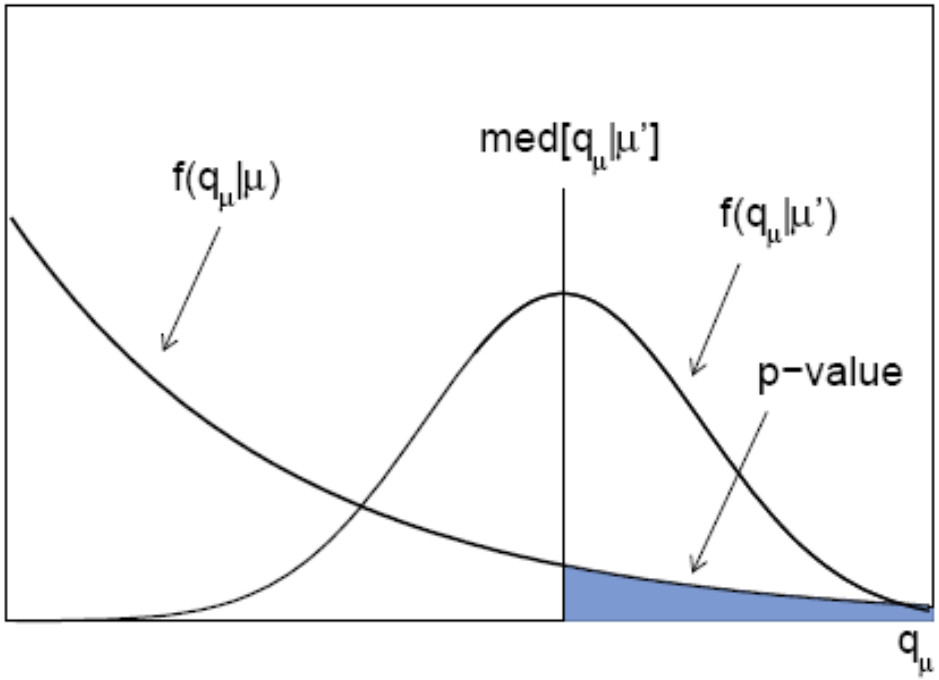
The dashed curve is the median μ_{up}, and the green (yellow) bands give the $\pm 1\sigma$ (2σ) regions of this distribution.

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



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 $f(q_0|\mu')$,

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $\dfrac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov: $\sqrt{2\left((s+b)\ln\left(1+\dfrac{s}{b}\right)-s\right)}$

II. Discovery sensitivity with uncertainty in b, σ_b:

(a) $\dfrac{s}{\sqrt{b+\sigma_b^2}}$

(b) Profile likelihood ratio test & Asimov:

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

Counting experiment with known background

Count a number of events $n \sim \text{Poisson}(s+b)$, 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,

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

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 $s + b$, $n \to x \sim$ 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 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:

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

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

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

Approximate Poisson significance (continued)

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

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

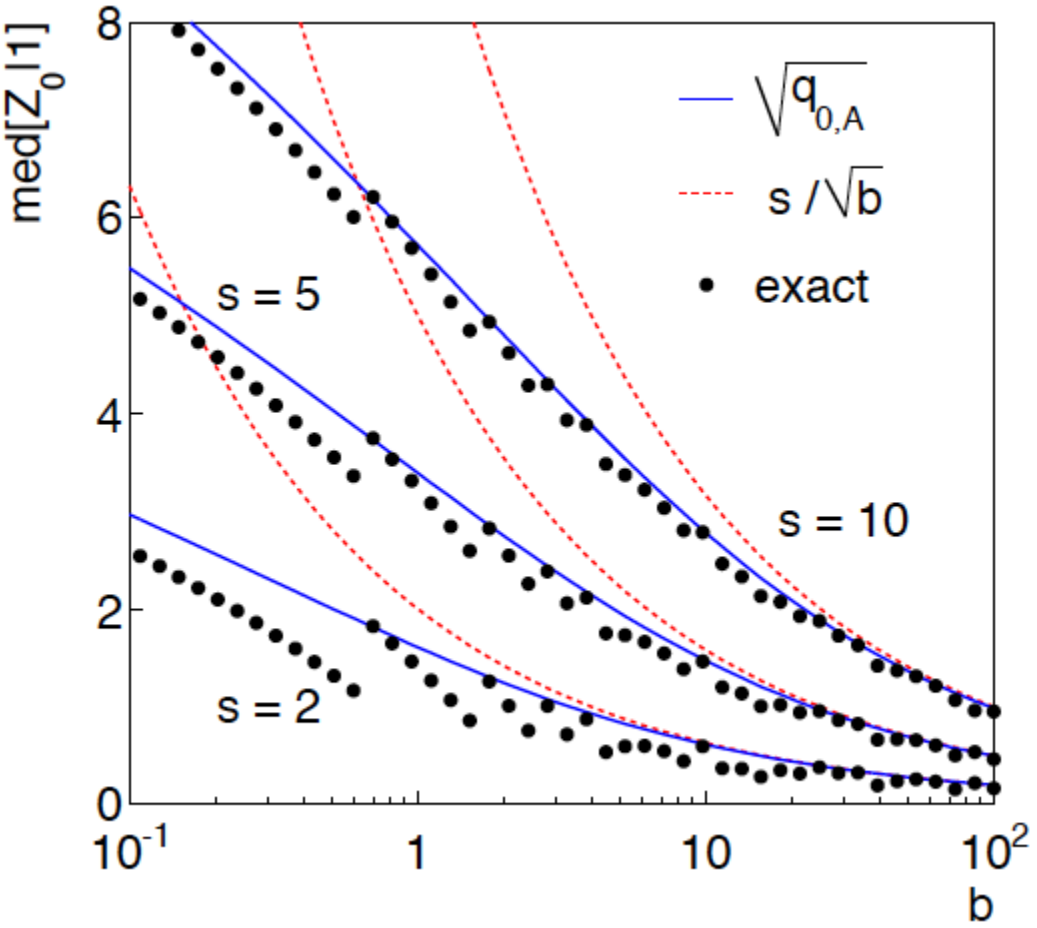
To find median$[Z|s]$, let $n \to s + b$ (i.e., the Asimov data set):

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

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

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

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

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

Profile likelihood with b uncertain

This is the well studied "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 parameter):

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

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau \,,$$

$$\hat{b} = m/\tau \,,$$

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

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:

$$n \to s + b$$

$$m \to \tau b$$

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

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

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

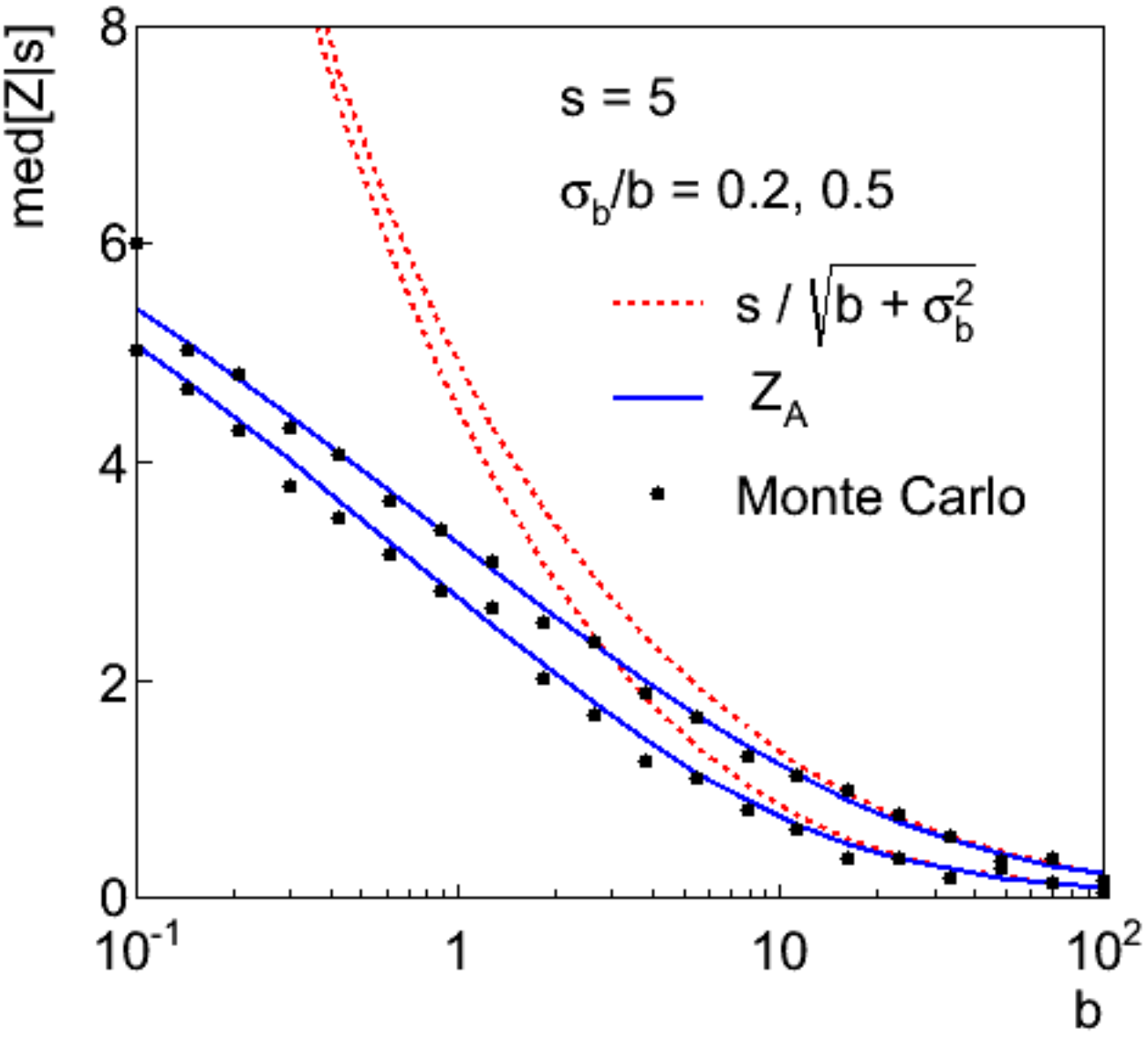
Limiting cases

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 the "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

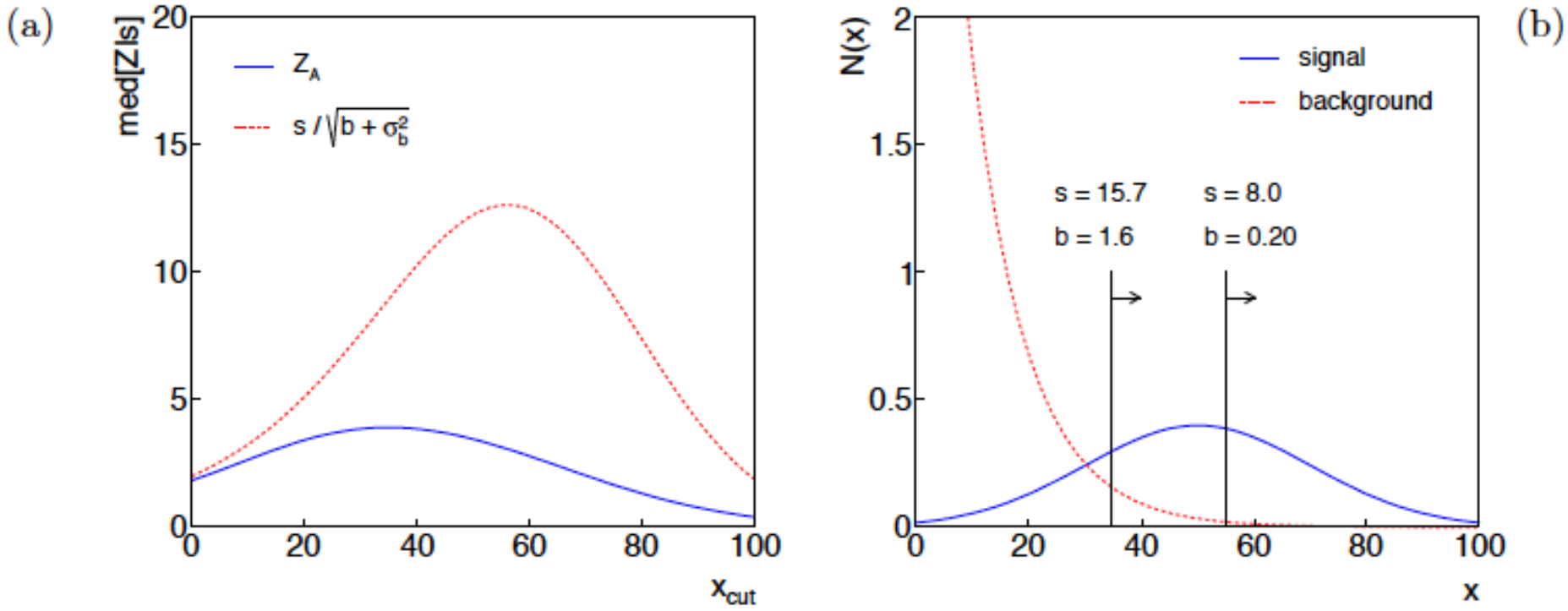


Figure 1: (a) The expected significance as a function of the cut value $x_{\rm cut}$; (b) the distributions of signal and background with the optimal cut value indicated.

Summary on discovery sensitivity

Simple formula for expected discovery significance based on profile likelihood ratio test and Asimov approximation:

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

For large b, all formulae OK.

For small b, $s/\sqrt{b}$ and $s/\sqrt{(b+\sigma_b^2)}$ overestimate the significance.

Could be important in optimization of searches with low background.

Formula maybe also OK if model is not simple on/off experiment, e.g., several background control measurements (checking this).

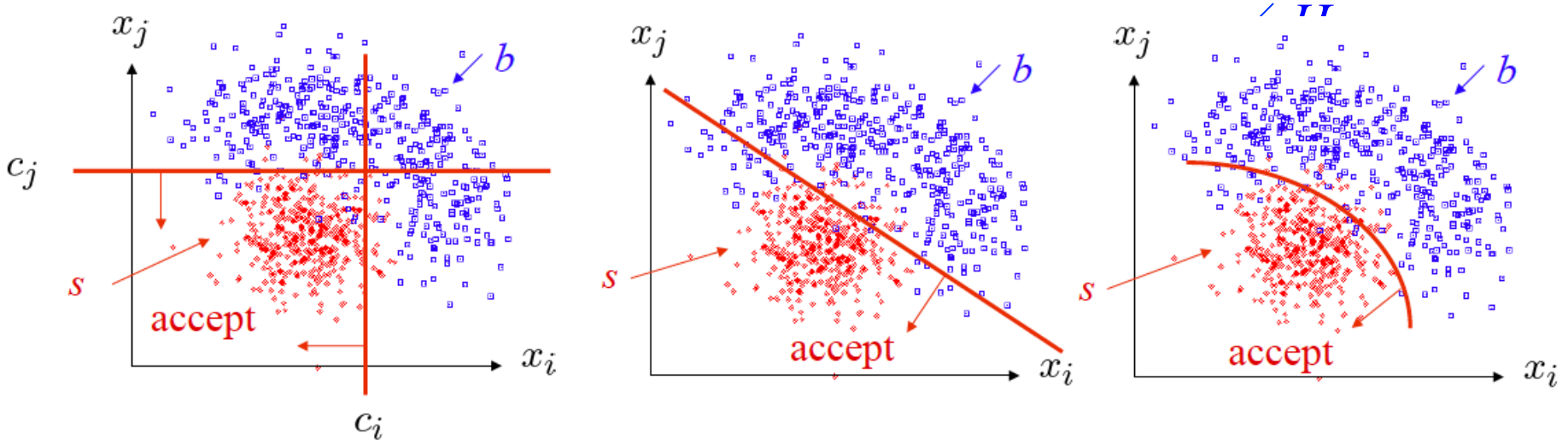
Prototype multivariate analysis in HEP

Each event yields a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons, x_2 = p_t of jet, ...

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

1) What kind of decision boundary best separates the two classes?



2) What is optimal test of hypothesis that event sample contains only background?

The Higgs Machine Learning Challenge

higgsml.lal.in2p3.fr

Competition ran summer 2014 on **kaggle.com**,

~2000 participants.

Many new ideas from machine learning community, currently being absorbed by HEP:

Deep learning
Cross validation
Ensemble methods
...

Comment on choice of variables for MVA

Usually when choosing the input variables for a multivariate analysis, one tries to find those that provide the most discrimination between the signal and background events.

But because of the correlations between variables, there are often variables that have identical distributions between signal and background, which nevertheless are helpful when used in an MVA.

A simple example is a variable related to the "quality" of an event, e.g., the number of pile-up vertices. This will have the same distribution for signal and background, but using it will allow the MVA to appropriately weight those events that are better measured and deweight (without completely dropping) the events where there is less information.

A simple example (2D)

Consider two variables, x_1 and x_2, and suppose we have formulas for the joint pdfs for both signal (s) and background (b) events (in real problems the formulas are usually not available).

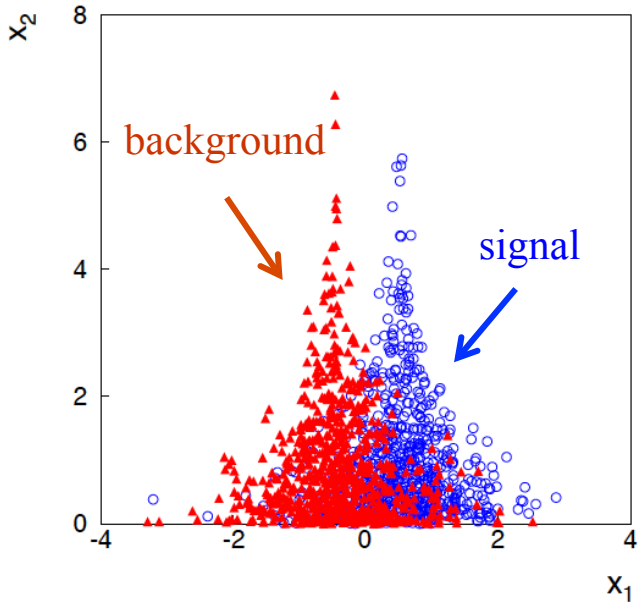
$f(x_1|x_2)$ ~ Gaussian, different means for s/b,
Gaussians have same σ, which depends on x_2,
$f(x_2)$ ~ exponential, same for both s and b,
$f(x_1, x_2) = f(x_1|x_2)\, f(x_2)$:

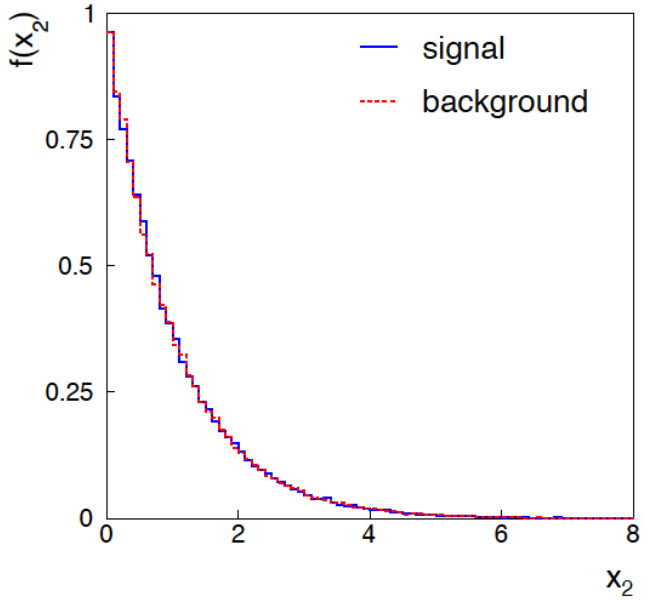
$$f(x_1, x_2|\mathrm{s}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{s})^2/2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

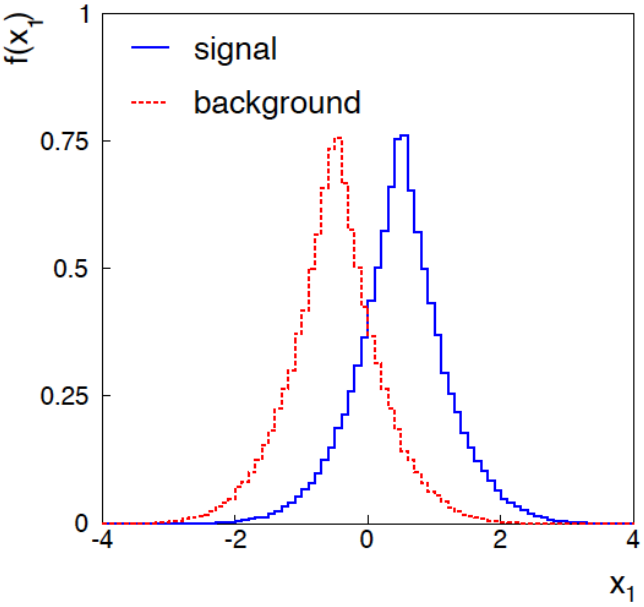
$$f(x_1, x_2|\mathrm{b}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{b})^2/2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$

$$\sigma(x_2) = \sigma_0 e^{-x_2/\xi}$$

Joint and marginal distributions of x_1, x_2







Distribution $f(x_2)$ same for s, b.

So does x_2 help discriminate between the two event types?

Likelihood ratio for 2D example

Neyman-Pearson lemma says best critical region is determined by the likelihood ratio:

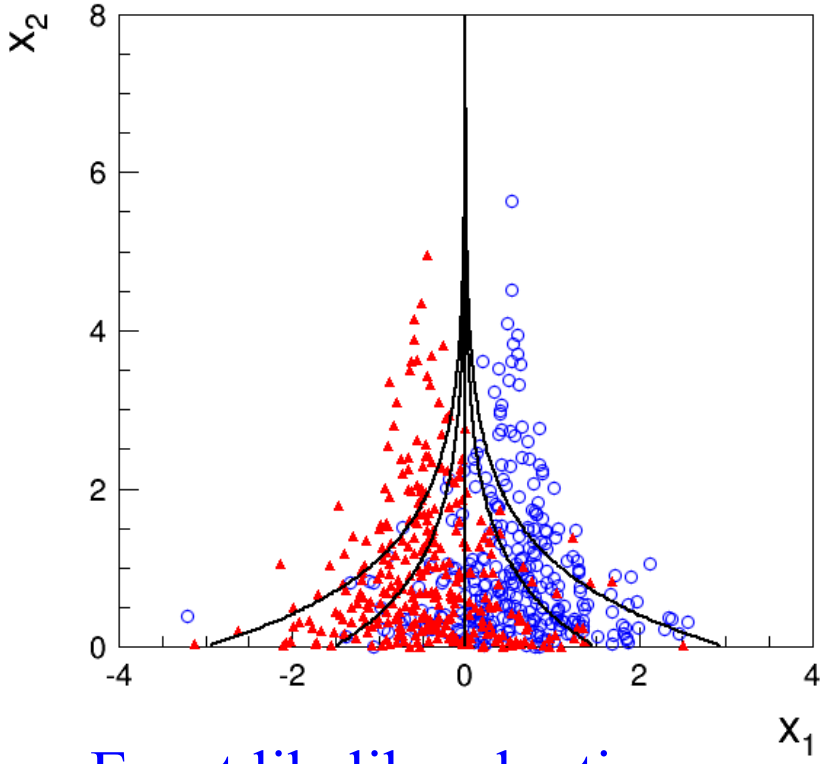
$$t(x_1, x_2) = \frac{f(x_1, x_2|\mathrm{s})}{f(x_1, x_2|\mathrm{b})}$$

Equivalently we can use any monotonic function of this as a test statistic, e.g.,

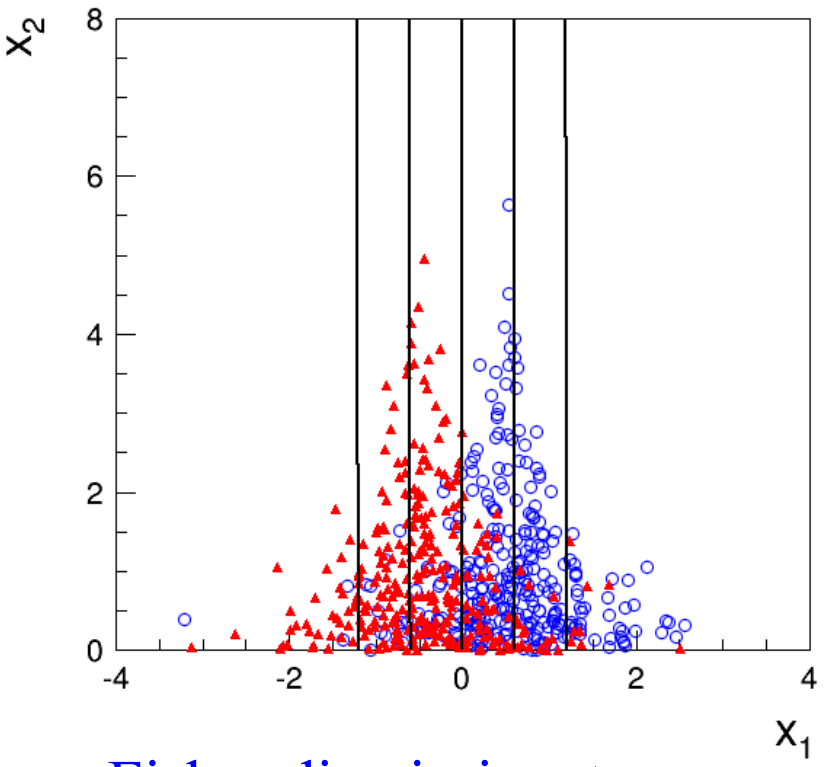
$$\ln t = \frac{\frac{1}{2}(\mu_\mathrm{b}^2 - \mu_\mathrm{s}^2) + (\mu_\mathrm{s} - \mu_\mathrm{b})x_1}{\sigma_0^2 e^{-2x_2/\xi}}$$

Boundary of optimal critical region will be curve of constant ln t, and this depends on x_2!

Contours of constant MVA output

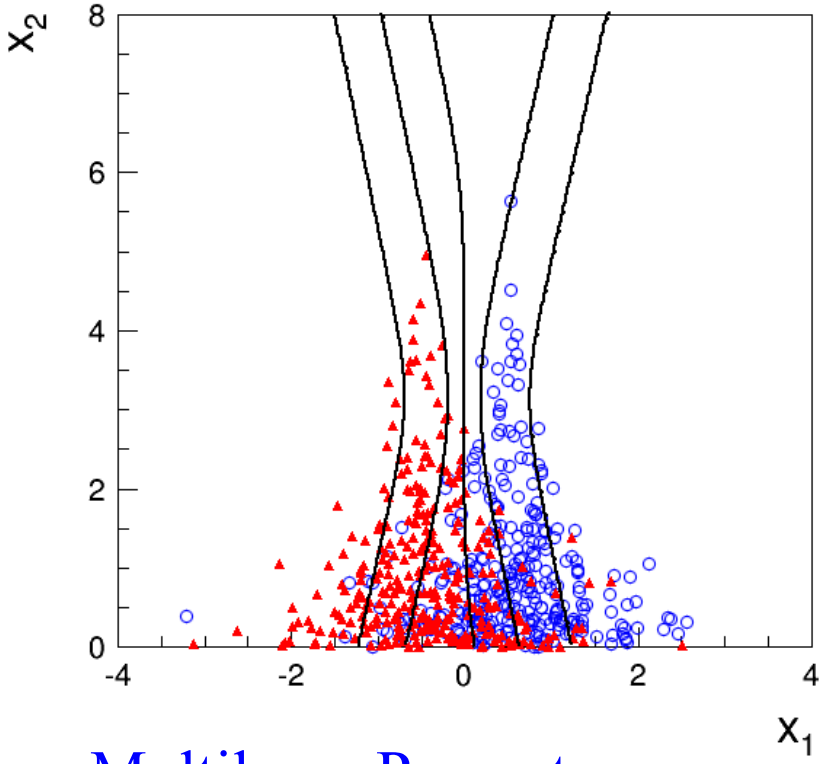


Exact likelihood ratio

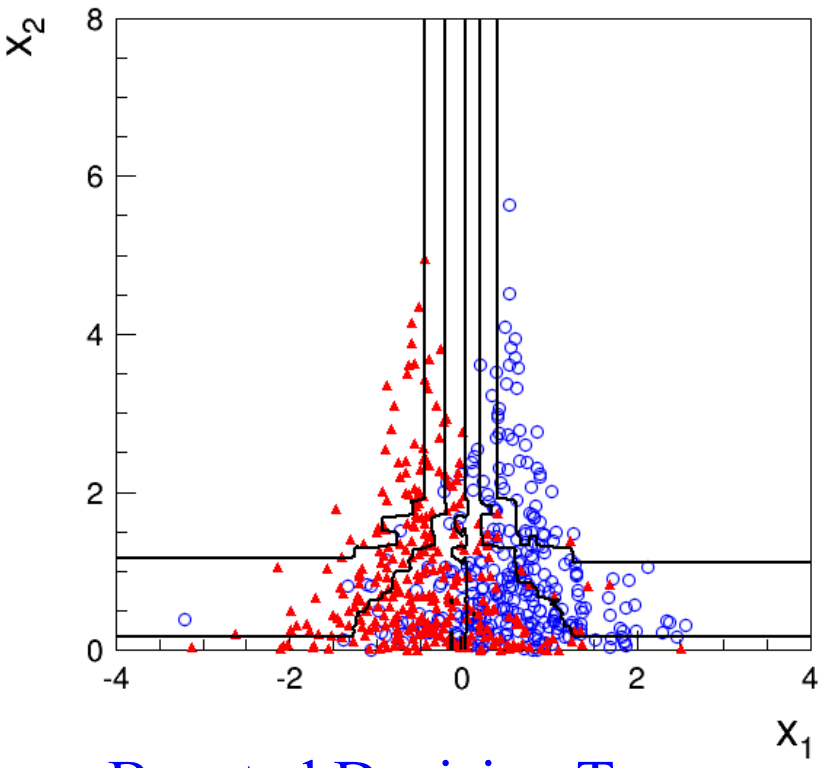


Fisher discriminant

Contours of constant MVA output



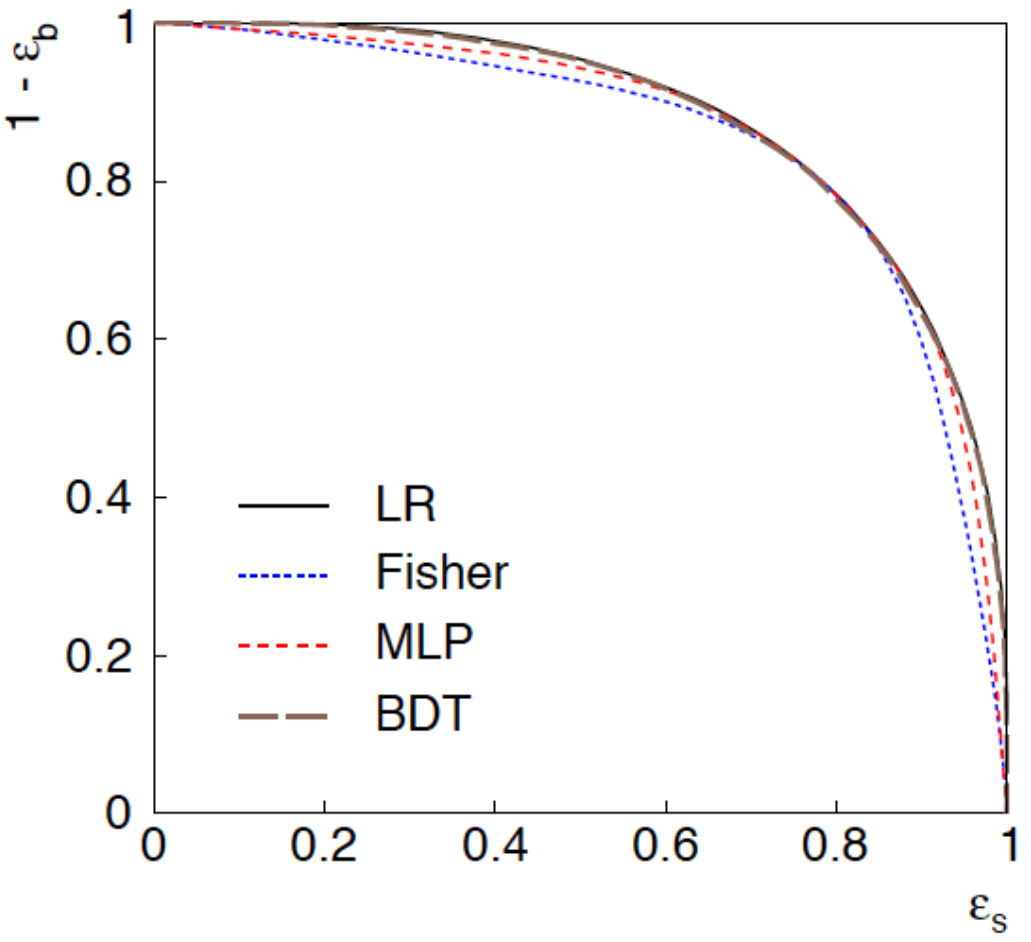
Multilayer Perceptron
1 hidden layer with 2 nodes



Boosted Decision Tree
200 iterations (AdaBoost)

Training samples: 10^5 signal and 10^5 background events

ROC curve



ROC = "receiver operating characteristic" (term from signal processing).

Shows (usually) background rejection $(1-\varepsilon_b)$ versus signal efficiency ε_s.

Higher curve is better; usually analysis focused on a small part of the curve.

2D Example: discussion

Even though the distribution of x_2 is same for signal and background, x_1 and x_2 are not independent, so using x_2 as an input variable helps.

Here we can understand why: high values of x_2 correspond to a smaller σ for the Gaussian of x_1. So high x_2 means that the value of x_1 was well measured.

If we don't consider x_2, then all of the x_1 measurements are lumped together. Those with large σ (low x_2) "pollute" the well measured events with low σ (high x_2).

Often in HEP there may be variables that are characteristic of how well measured an event is (region of detector, number of pile-up vertices,...). Including these variables in a multivariate analysis preserves the information carried by the well-measured events, leading to improved performance.

Summary and conclusions

Statistical methods continue to play a crucial role in HEP analyses; recent Higgs discovery is an important example.

HEP has focused on frequentist tests for both p-values and limits; many tools developed, e.g.,

asymptotic distributions of tests statistics, (CCGV arXiv:1007.1727, Eur Phys. J C 71(2011) 1544; recent extension (CCGV) in arXiv:1210:6948),

increasing use of advanced multivariate methods,...

Many other questions untouched today, e.g.,

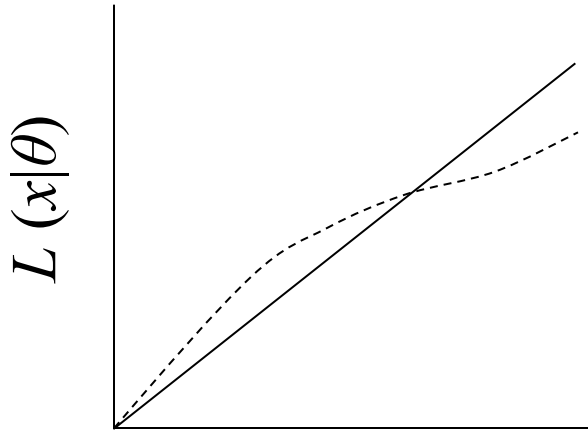
simple corrections for Look-Elsewhere Effect,

Use of Bayesian methods for both limits and discovery

Extra slides

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.

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

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).

Large sample distribution of the profile likelihood ratio (Wilks' theorem, cont.)

Suppose problem has likelihood $L(\boldsymbol{\theta}, \boldsymbol{\nu})$, with

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \qquad \leftarrow \text{parameters of interest}$$

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_M) \qquad \leftarrow \text{nuisance parameters}$$

Want to test point in $\boldsymbol{\theta}$-space. Define profile likelihood ratio:

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}, \hat{\hat{\boldsymbol{\nu}}}(\boldsymbol{\theta}))}{L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\nu}})} \quad , \quad \text{where} \quad \hat{\hat{\boldsymbol{\nu}}}(\boldsymbol{\theta}) = \underset{\boldsymbol{\nu}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}, \boldsymbol{\nu})$$

"profiled" values of $\boldsymbol{\nu}$

and define $q_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$.

Wilks' theorem says that distribution $f(q_{\boldsymbol{\theta}}|\boldsymbol{\theta},\boldsymbol{\nu})$ approaches the chi-square pdf for N degrees of freedom for large sample (and regularity conditions), independent of the nuisance parameters $\boldsymbol{\nu}$.

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

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

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 for reasons external to the analysis (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 profiled values of the nuisance parameters:

$\hat{\hat{\nu}}(\theta)$ "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 $(\theta, \hat{\hat{\nu}}(\theta))$.

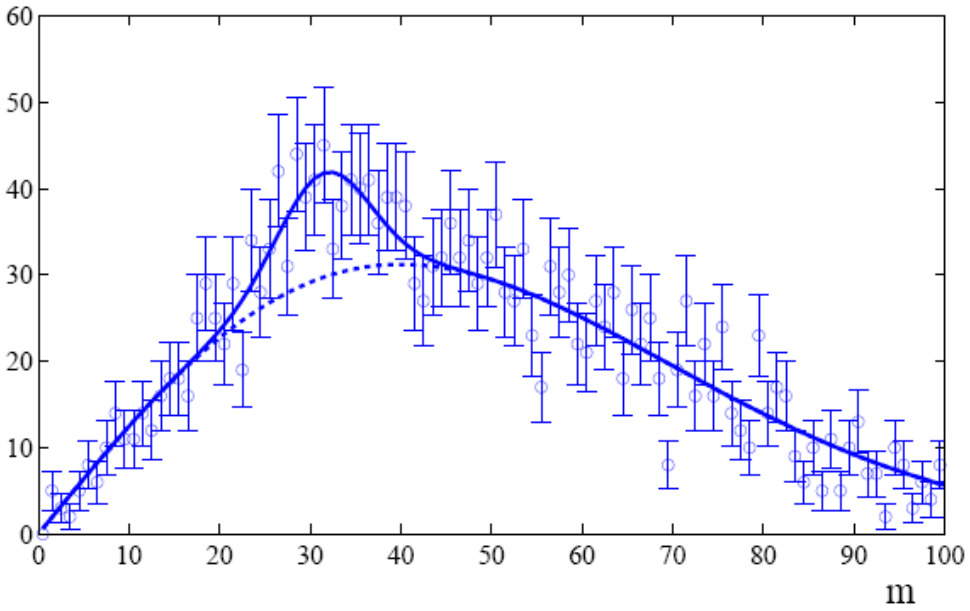
Elsewhere it may under- or overcover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

Gross and Vitells, EPJC 70:525-530,2010, arXiv:1005.1891

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

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

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

The resulting p-value

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

gives the probability to find a value of t_{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

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

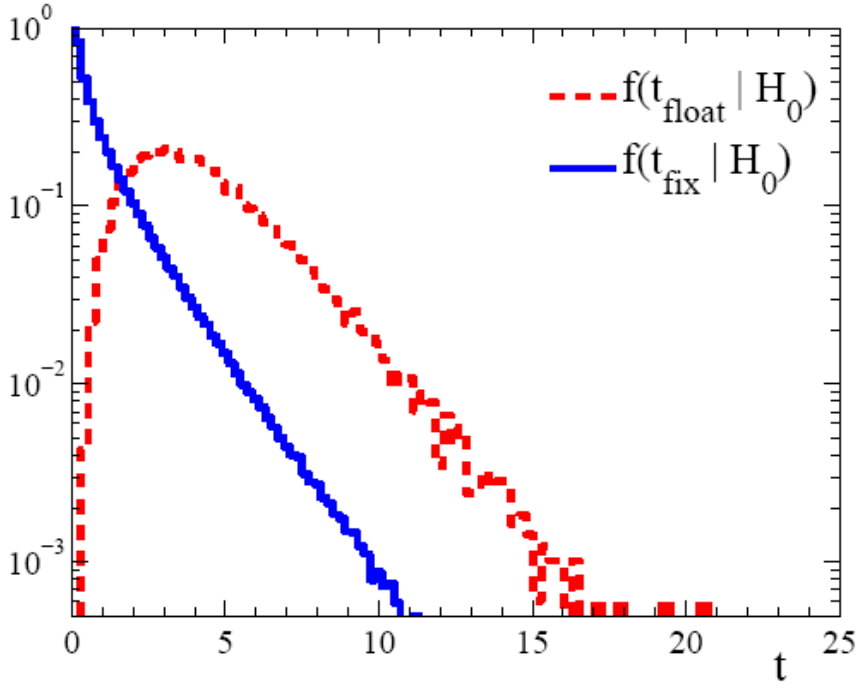
(Note m does not appear in the $\mu = 0$ model.)

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

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

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

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).

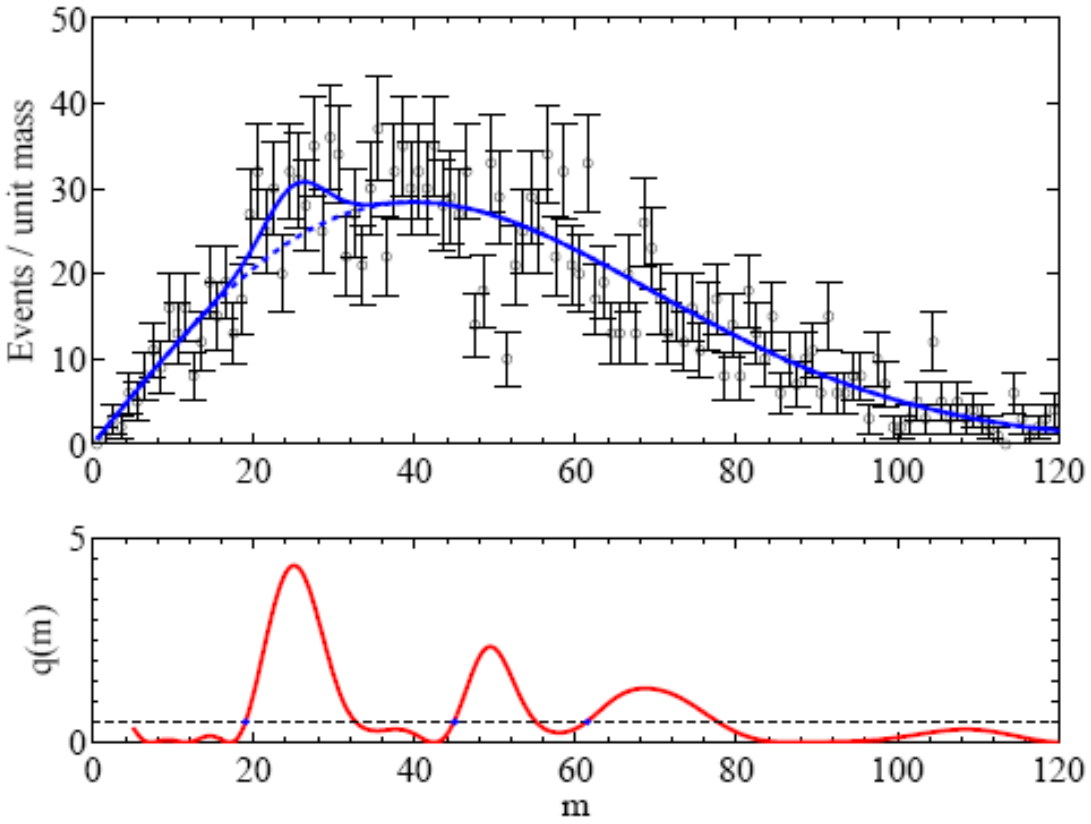
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires $\langle N(c)\rangle$, the mean number "upcrossings" of $t_{\rm fix} = -2\ln\lambda$ above a threshold $c = t_{\rm fix,obs}$ found when varying the mass m_0 over the range considered.

$\langle N(c)\rangle$ 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.



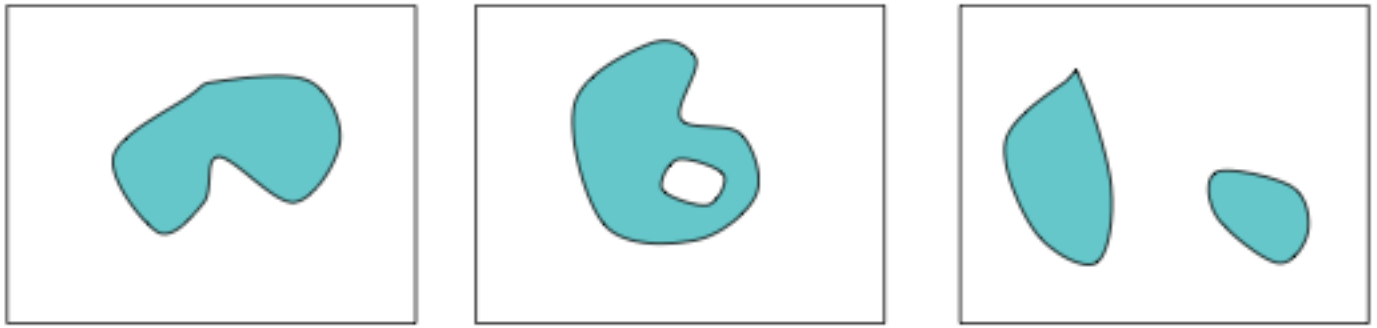
Vitells and Gross, Astropart. Phys. 35 (2011) 230-234; arXiv:1105.4355

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.