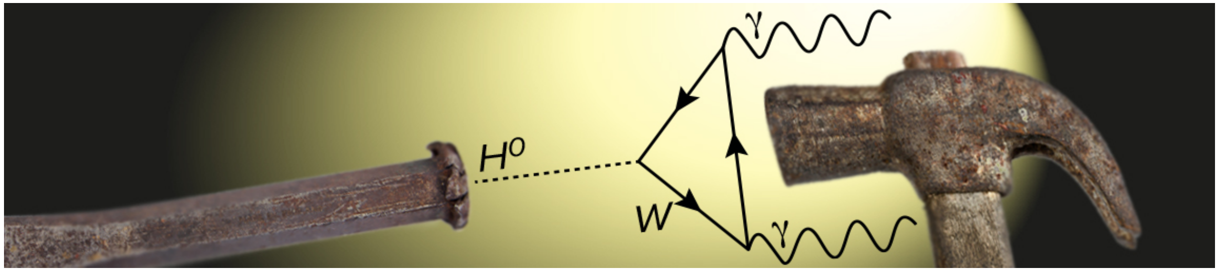
Statistical Treatment of Data in HEP



Hammers & Nails Workshop
Weizmann Institute of Science, Israel
20 July 2017

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

Plan for the next hour

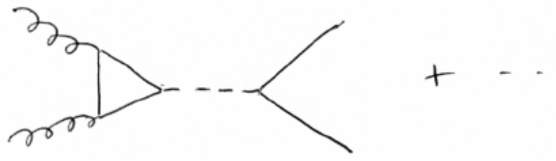
Goal of talk is to try to explain, mainly to the non-physics people, the statistical procedures usually used in HEP.

Will focus on statistical tests used e.g. for discovering a new signal process, and how this relates to event classification.

Theory ↔ Statistics ↔ Experiment

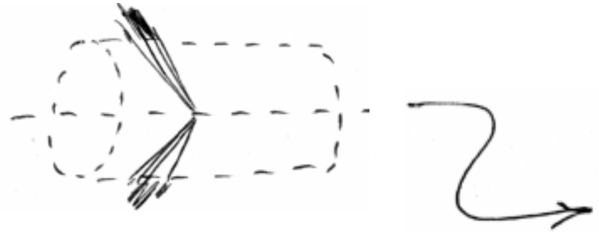
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + \cdots$$

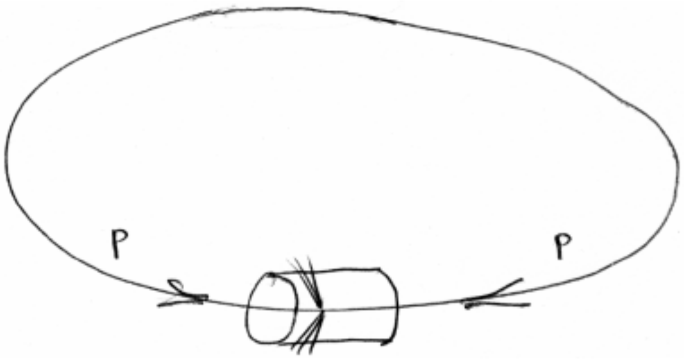


$$\sigma = \frac{G_F \, \alpha_s^2 \, m_H^2}{288 \sqrt{2\pi}} \times \sim$$

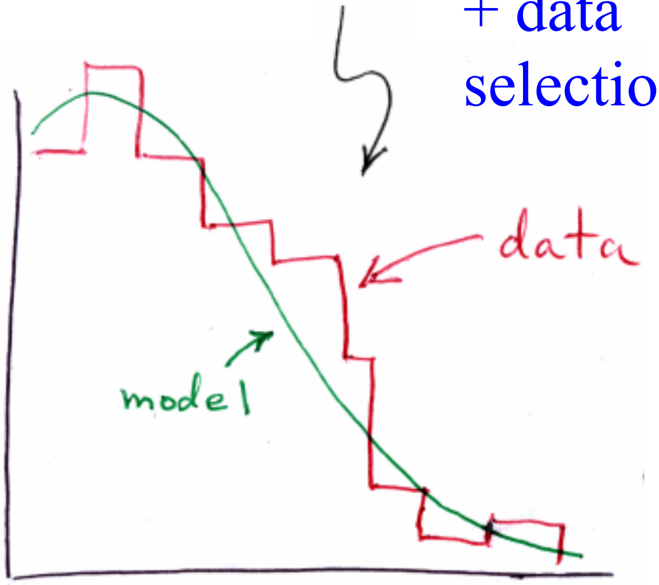
+ simulation of detector and cuts



Experiment:



+ data selection



Goals of statistical data analysis in HEP

Start with probabilistic model for data, usually with parameters.

Estimate the parameters; characterize accuracy.

Test hypotheses: Reject SM → "discovery" of New Physics

Out of the two main interpretations of probability,

Frequentist – A = outcome of repeatable experiment:

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

Subjective (Bayesian) – A = hypothesis (true or false):

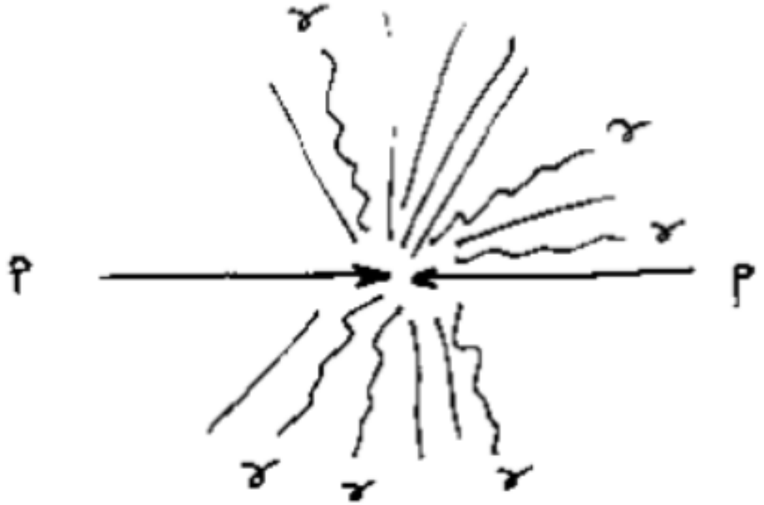
$$P(A) = \text{degree of belief that } A \text{ is true}$$

HEP uses mainly frequentist methods (will focus on this today).

The Data

The data $\boldsymbol{x}$ = raw (10^8 numbers), reduced (~10-100 numbers),
or highly reduced (single number)

$\boldsymbol{x}$ characterizes: an individual particle,
or the outcome of a particle collision (an "event"),
or an entire dataset (a set of events)



E.g. for case of an event, could have $\boldsymbol{x} = (x_1, x_2, ..., x_n)$,
x_1 = number of jets, x_2 = energy of highest energy jet, ... etc.

The Model(s)

Probability to observe the data given by hypothesis (model) H:

$$P(\boldsymbol{x}|H)$$

Often composite:

$$P(\boldsymbol{x}|H, \boldsymbol{\theta})$$

$\boldsymbol{\theta}$ contains in general parameter(s) of interest (e.g., the rate of an new signal process) and nuisance parameters (e.g., rates of background processes, calibration constants).

If viewed as function of H, $\boldsymbol{\theta}$, then this is the likelihood:

$$L(H, \boldsymbol{\theta}) = P(\boldsymbol{x}|H, \boldsymbol{\theta})$$

The “Standard Model” (19 parameters, mostly well determined)

Some alternatives: supersymmetry, extra dimensions, ...

(Part of) the Standard Model

$$\begin{aligned}\mathcal{L} =& \sum_i \bar{\psi}_i \left(i \not\partial - m_i - \frac{g m_i H}{2M_W} \right) \psi_i - \frac{g}{2\sqrt{2}} \sum_i \bar{\Psi}_i \gamma^\mu (1-\gamma^5)(T^+ W^+_\mu + T^- W^-_\mu) \Psi_i \\ & - e \sum_i q_i \bar{\psi}_i \gamma^\mu \psi_i A_\mu - \frac{g}{2\cos\theta_W} \sum_i \bar{\psi}_i \gamma^\mu (g_V^i - g_A^i \gamma^5) \psi_i Z_\mu \\ & + \sum_q \bar{\psi}_{q,a}(i\gamma^\mu \partial_\mu \delta_{ab} - g_s \gamma^\mu t^C_{ab} A^C_\mu - m_q \delta_{ab})\psi_{q,b} - \frac{1}{4} F^A_{\mu\nu} F^{A\,\mu\nu} \quad + \ldots\end{aligned}$$

In principle this determines $P(\boldsymbol{x}|\boldsymbol{\theta}_{\mathrm{SM}})$.

In practice only have Monte Carlo models, i.e., we can generate events $\boldsymbol{x} \sim P(\boldsymbol{x}|\mathrm{SM})$,

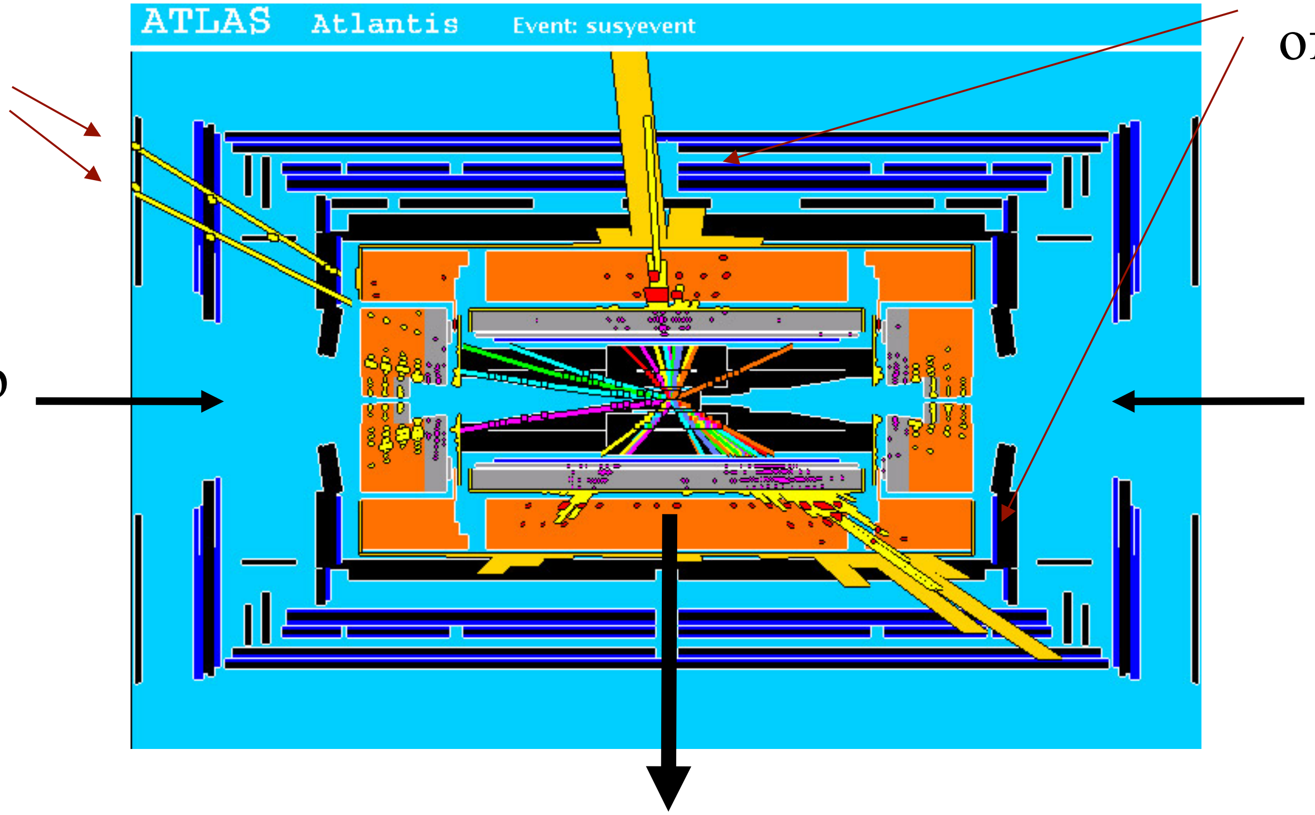
Also for a variety of alternatives: $\boldsymbol{x} \sim \mathrm{P}(\boldsymbol{x}|\mathrm{SUSY})$, etc.

Even then many approximations needed (perturbation theory, nonperturbative effects, approximate detector response,..)

A simulated SUSY event in ATLAS

high p_T muons

high p_T jets of hadrons

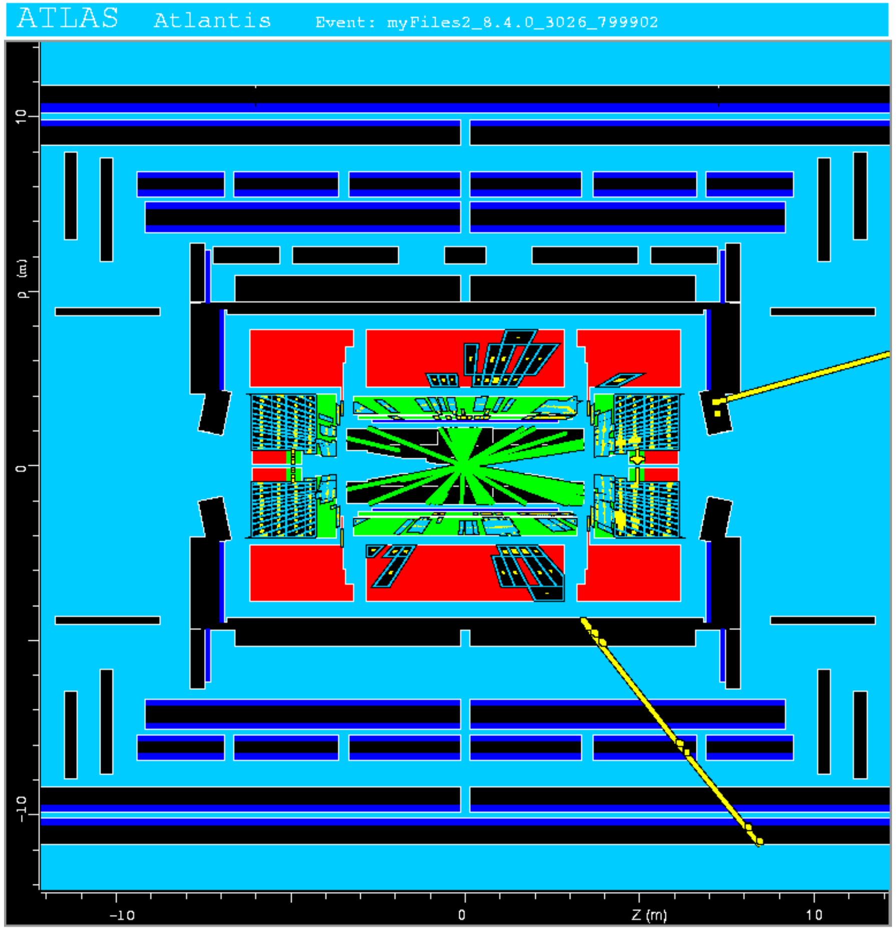


p

p

missing transverse energy

A top-antitop (Standard Model) event



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

→ can easily mimic a SUSY event.

Search for presence of "New Physics" with a Statistical Test.

Frequentist Statistical Tests

Consider

data $\boldsymbol{x}$,

model to test (the null) $P(\boldsymbol{x}|H_0)$,

an alternative model $P(\boldsymbol{x}|H_1)$.

Define critical region w such that for a given (small) size α

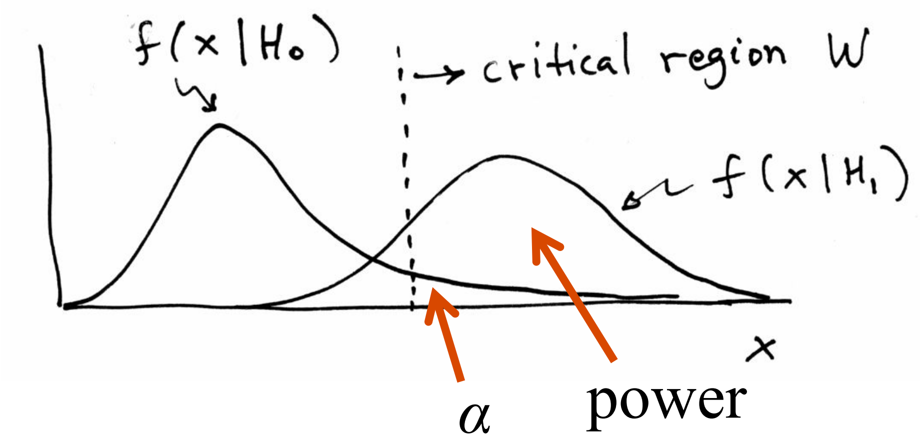
$$P(\boldsymbol{x} \in w|H_0) \leq \alpha$$

Choose critical region to maximimize power M with respect to H_1

$$M(H_1) = P(\boldsymbol{x} \in w|H_1)$$

Do the measurement.

If $\boldsymbol{x} \in w$, reject H_0.



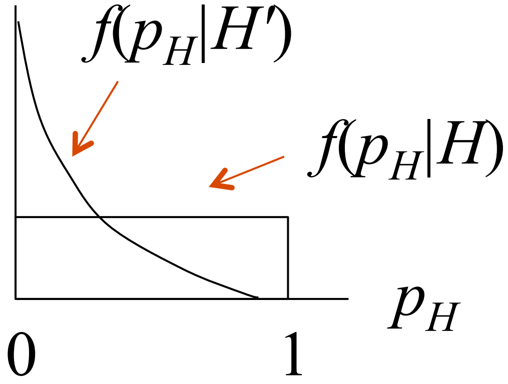
p-values

Often formulate test in terms of p-value:

$$p_H = P(\boldsymbol{x} \in \text{region of equal or lesser compatibility} \mid H)$$

"Less compatible with H" means "more compatible with alt. H'"

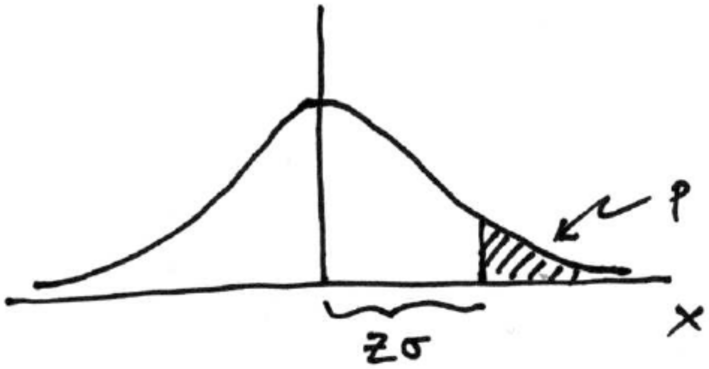
Distribution $f(p_H|H)$ uniform on [0,1], so can define critical region of a test as the region where the p-value is $\leq \alpha$.



Formally the p-value relates only to H but the resulting test will have a given power with respect to a given alternative 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
```

E.g. $Z = 5$ (a "5 sigma effect") corresponds to $p = 2.9 \times 10^{-7}$.

Constructing an optimal test

Neyman-Pearson lemma:

When choosing critical region w of test of H_0 of a given size α, to obtain highest power with respect to H_1, w should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$$

inside the region, and $< c_\alpha$ outside, where c_α is a constant chosen to give a test of the desired size.

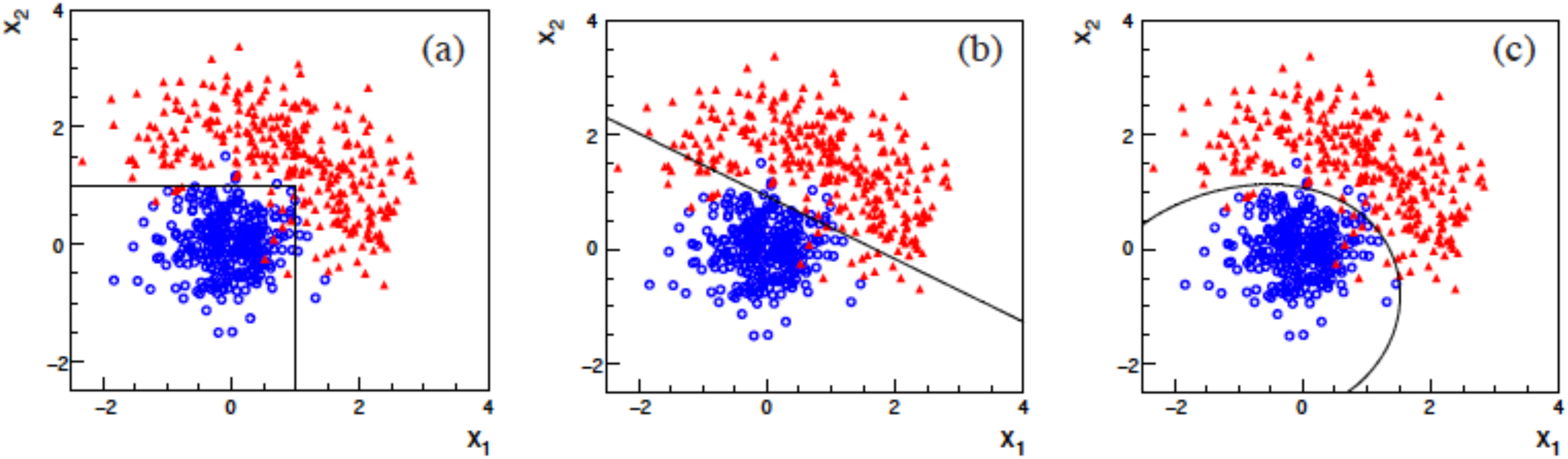
Equivalently, optimal scalar test statistic is the likelihood ratio

$$r(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

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

Event Classification viewed as a test

Suppose signal (s) and background (b) events have data $\boldsymbol{x}$ that follow $p(\boldsymbol{x}|s)$, $p(\boldsymbol{x}|b)$. From simulated data find:



Can test for each event hypothesis that it is of type b.

Best critical region = ? ("cuts", linear, nonlinear,...)

Define statistic $y(\boldsymbol{x})$ such that boundary of critical region is $y(\boldsymbol{x}) = y_c$, using e.g., neural network, BDT, ..., optimally something that is a monotonic function of $r(\boldsymbol{x}) = p(\boldsymbol{x}|s) / p(\boldsymbol{x}|b)$.

Test for discovery of signal process

Goal: search for events from an undiscovered signal process s in a sample of events otherwise consisting of background b.

Measure $\boldsymbol{x}$ for each event: $\boldsymbol{x} \sim p(\boldsymbol{x}|s)$ or $p(\boldsymbol{x}|b)$ (only have generative models, no closed formulae).

Suppose we observe n events, data consist of: n, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$,

Goal is to test H_0 : all events are of background type b

versus H_1 : event sample contains some events of signal type s

Suppose number of events $n \sim \text{Poisson}(\mu s + b)$, where here s, b = expected number of events of corresponding type, (assume approx. known) and μ = signal strength parameter, i.e.,

$$H_0 \text{ means } \mu = 0, \quad H_1 \text{ (usually) means } \mu > 0.$$

Optimal test for discovery

Likelihood function is:

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \prod_{i=1}^{n} \left[\frac{\mu s}{\mu s + b} p(\mathbf{x}_i | s) + \frac{b}{\mu s + b} p(\mathbf{x}_i | b) \right]$$

Neyman-Pearson say optimal statistic for test of $\mu = 0$ versus alternative of nonzero μ is

$$\frac{L(\mu)}{L(0)} = e^{-\mu s} \prod_{i=1}^{n} \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i | s)}{p(\mathbf{x}_i | b)} \right)$$

or take log and drop constant term $-\mu s$,

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i | s)}{p(\mathbf{x}_i | b)} \right)$$

Relation to optimal event classifier

Optimal event classifier is (monotonic function of)

$$r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)}$$

But the ratio of distributions of *r* obeys

$$\frac{p(r|s)}{p(r|b)} = r(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|b)}$$

For a monotonic function *y*(*r*), *s* and *b* pdfs transform with same Jacobian, so

$$\frac{p(y|s)}{p(y|b)} = \frac{p(r|s)}{p(r|b)}$$

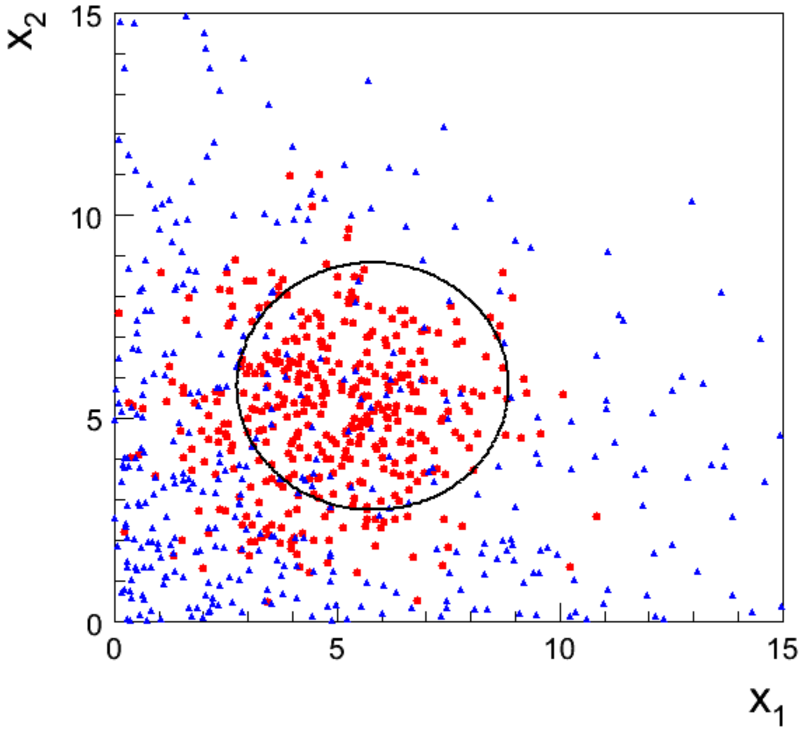
The statistic *Q* becomes (same as before!)

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(y_i|s)}{p(y_i|b)} \right)$$

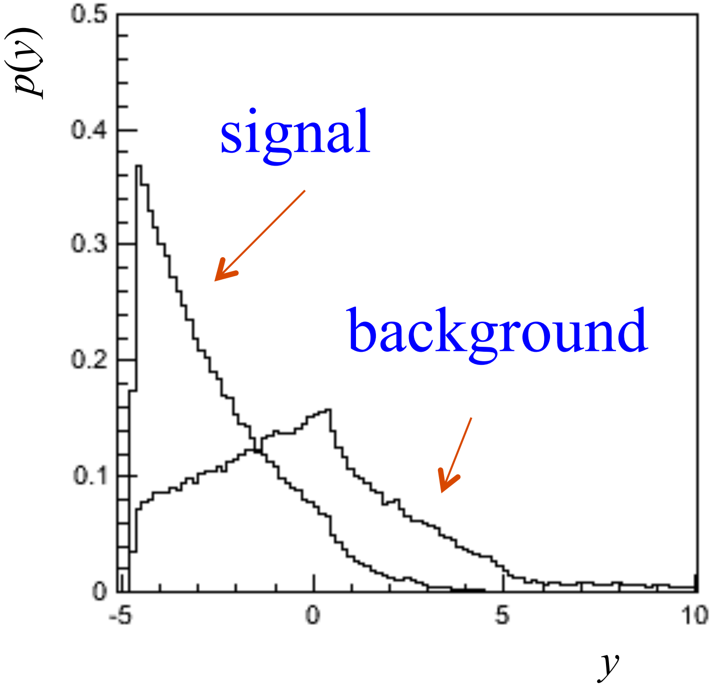
So if we find an event classifier *y*(***x***) that is a monotonic function of the (optimal) LR, and then use Monte Carlo models to determine, the pdfs ~ *p*(*y*|*s*) and *p*(*y*|*b*), then we can get the optimal *Q* to test whole sample for presence of signal.

Kyle Cranmer, Juan Pavez, Gilles Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, eprint: arXiv:1506.02169 [stat.AP] (2015).

Toy example



signal (red): $p(\boldsymbol{x}|s)$,
background (blue): $p(\boldsymbol{x}|\mathrm{b})$,
and contour of constant ratio



signal

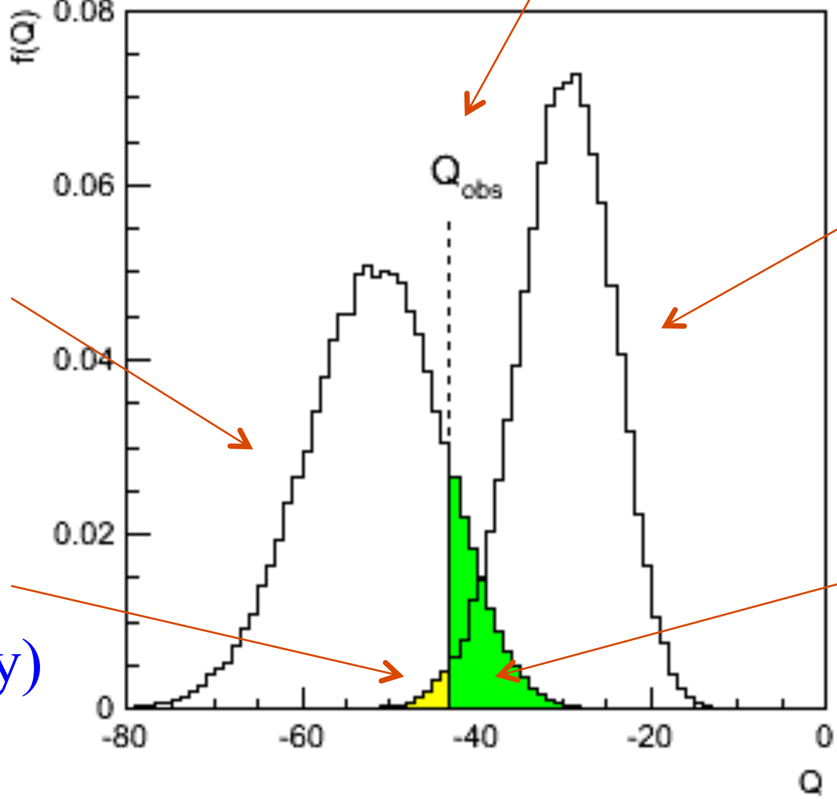
background

Distribution of event classifier
$y = -2 \ln [p(\boldsymbol{x}|s)/p(\boldsymbol{x}|\mathrm{b})]$

Distribution of Q

Suppose in real experiment Q is observed here.

$\mu = 1$



$f(Q|\mu = 1)$

$f(Q|\mu = 0)$

p-value of $\mu = 0$ (background only)

p-value of $\mu = 1$ (signal plus background)

If $p_\mu < \alpha$, reject signal model μ at confidence level $1 - \alpha$.

If $p_0 < 2.9 \times 10^{-7}$, reject background-only model (signif. $Z = 5$).

Parameter estimation

Most commonly used estimator of a a parameter θ from Maximum Likelihood:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(x|\theta)$$

Usually get covariance of estimators from 2nd derivatives of log-likelihood:

$$V_{ij} = \operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

$$V_{ij}^{-1} \approx -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

In general they may have a nonzero bias:

$$b = E[\hat{\theta}] - \theta$$

Least Squares used if measurements approx. Gaussian (and then equivalent to Maximum Likelihood) e.g. for tracking problems.

ML/LS estimator may not in some cases be regarded as the optimal trade-off between bias/variance e.g. in problems with large numbers of poorly constrained parameters (cf. regularized unfolding).

Interval estimation, limits

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

The values that are not rejected constitute a *confidence interval* for μ at confidence level CL = 1 – α.

> The confidence interval will by construction contain the true value of μ with probability of at least 1 – α.

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

Equivalently, the parameter values in the confidence interval have p-values of at least α.

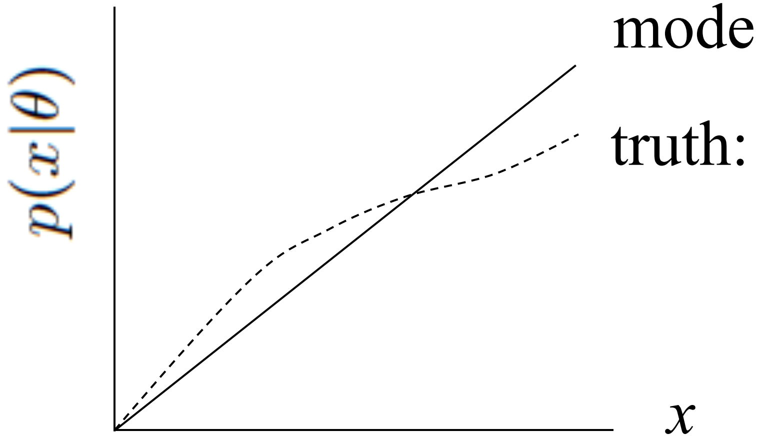
To find edge of interval (the "limit"), set $p_\mu = \alpha$ and solve for μ.

If null intervals, i.e., $p_\mu < \alpha$ for all μ, $p_\mu \to \frac{p_\mu}{1 - p_0}$ ("CL$_s$")

or use two-sided LR test (Feldman-Cousins / unified intervals)

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



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

truth: $p(x|\theta, \boldsymbol{\nu}) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$p(x|\theta) \rightarrow p(x|\theta, \boldsymbol{\nu})$$

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).

p-values in cases with nuisance parameters

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

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

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

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

→ “exact” confidence interval

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

But in general for finite data samples this is not true; one may be unable to reject some θ values if all values of ν must be considered (resulting interval for θ “overcovers”).

Profile construction (“hybrid resampling”)

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

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

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

“double hat” notation means profiled value, i.e., 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).

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

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

signal

background

Prototype analysis (II)

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

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

Assume the m_i are Poisson distributed with expectation values

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

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

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

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes independent of nuisance parameters in large sample limit.

Test statistic for discovery

Try to reject background-only (μ= 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.

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

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.

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

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

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

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

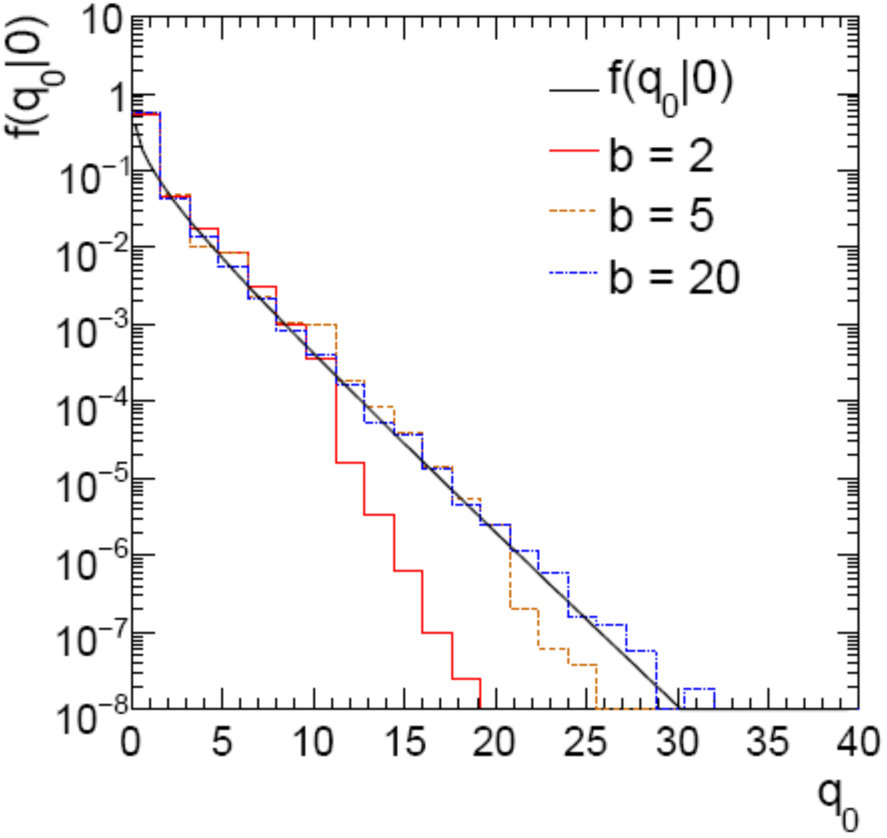
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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

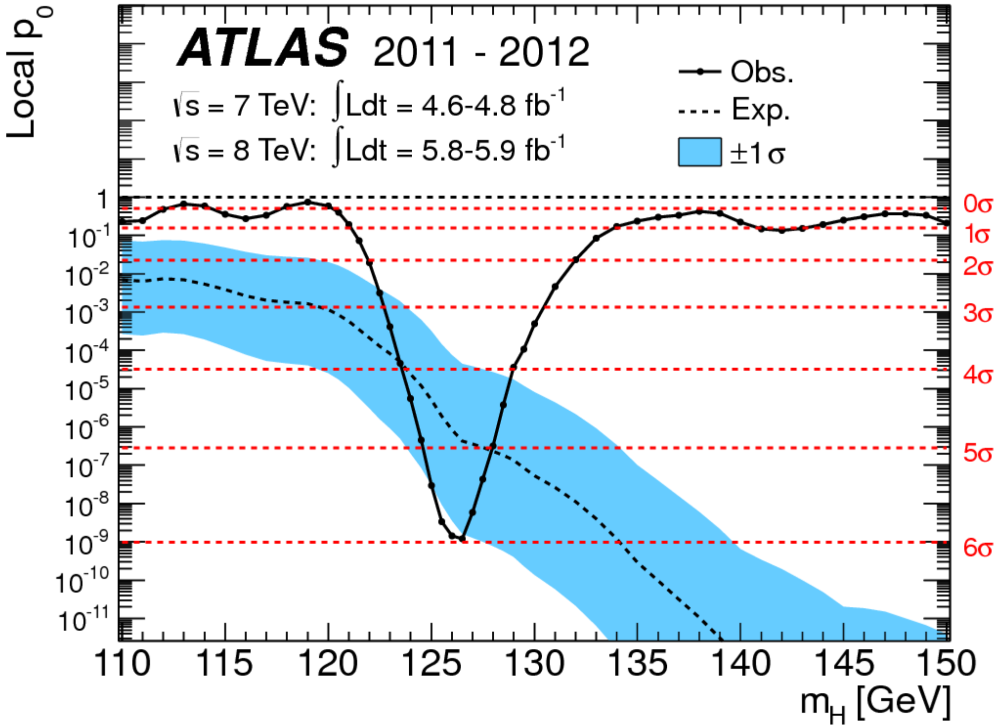


Discovery: the p_0 plot

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 (±1σ) of significances under assumption of the SM Higgs.

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

Test statistic for upper limits

For purposes of setting an upper limit on *μ* use

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

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

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

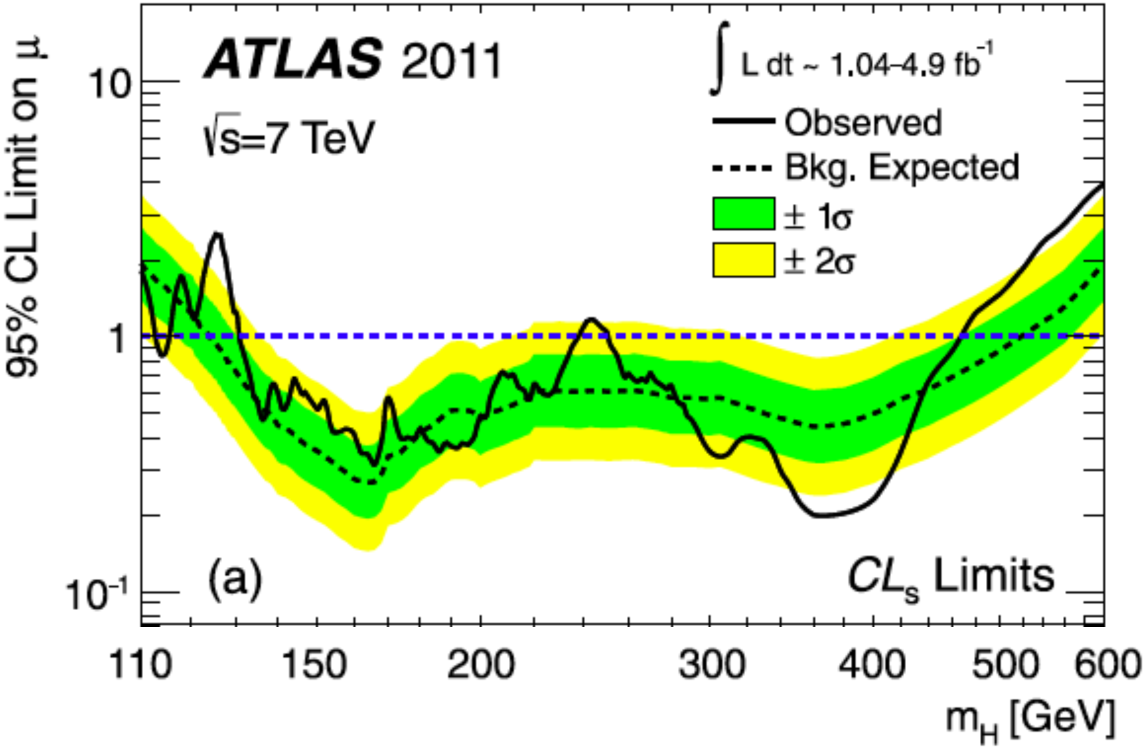
95% CL upper limit on *μ* is highest value for which *p*-value is not less than 0.05.

Example of upper limit on strength parameter μ

For every value of m_H, find the CLs upper limit on μ (solid line)

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

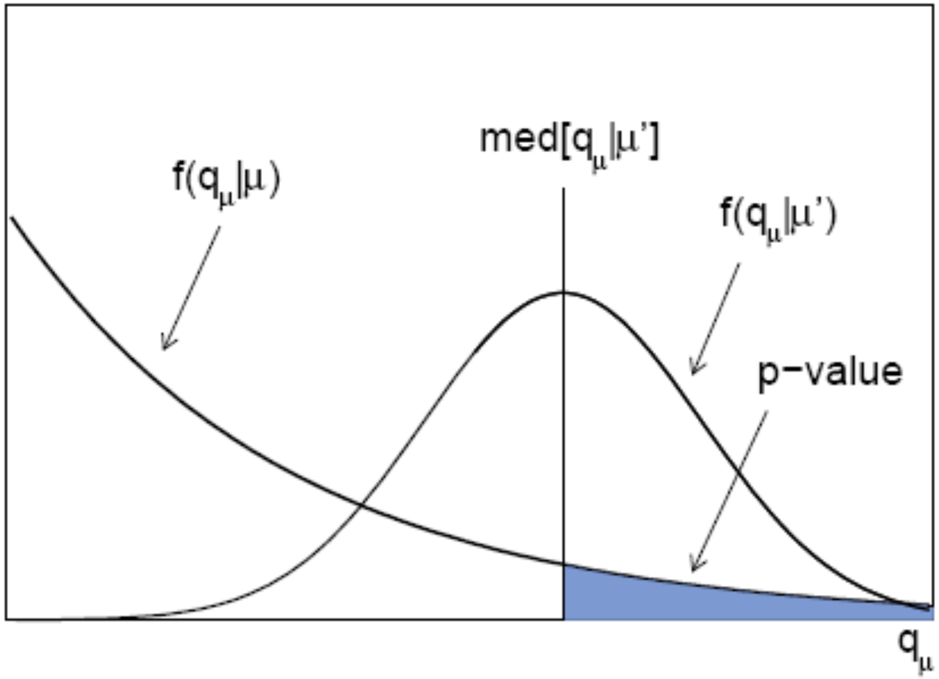
The dashed curve is the median μ_{up}, and the green (yellow) bands give the ± 1σ (2σ) regions of this distribution.



ATLAS, Phys. Lett. B 710 (2012) 49-66

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) $$\frac{s}{\sqrt{b}}$$

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

II. Discovery sensitivity with uncertainty in *b*, σ_b:

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

(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 $n \sim$ 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,

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

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

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

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

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

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

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

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 significance

Poisson likelihood for parameter *s* is

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

For now no nuisance params.

To test for discovery use profile likelihood ratio:

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

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

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

Approximate Poisson significance (continued)

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

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

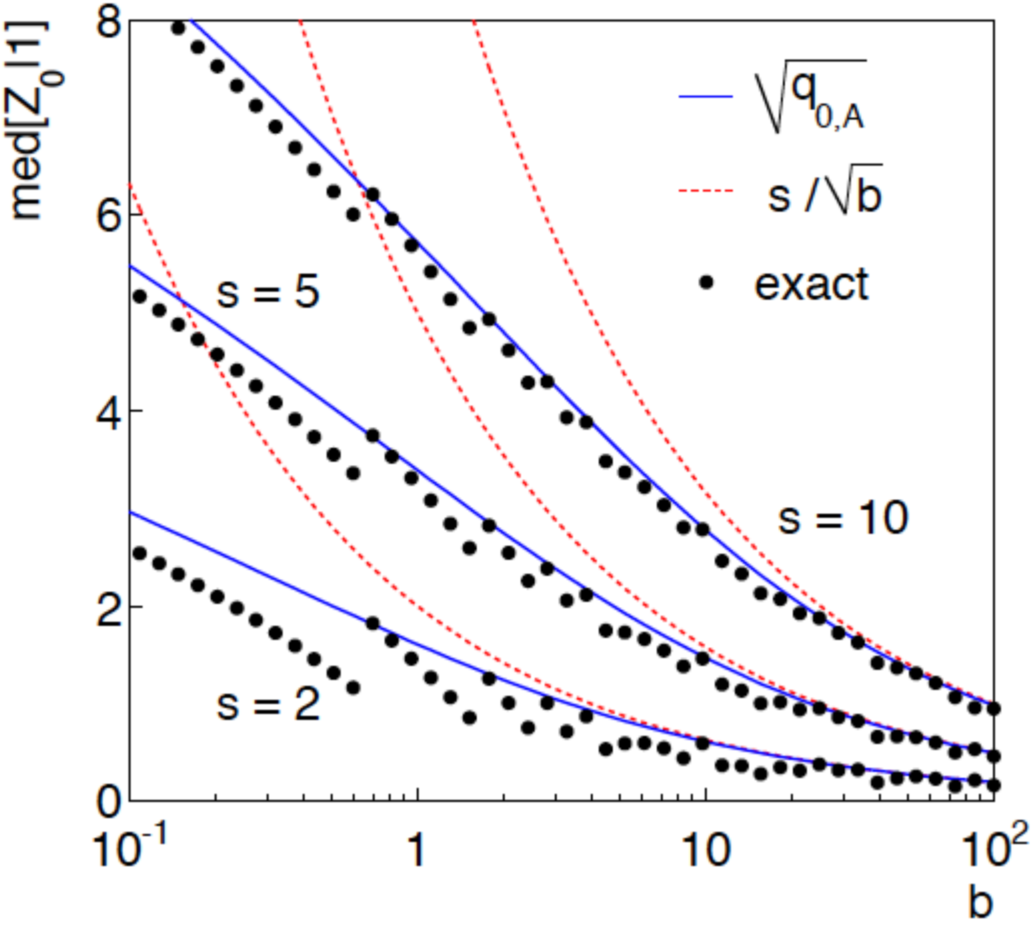
To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

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

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

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

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



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

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

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

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

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

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

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

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

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

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 parmeter):

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

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

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

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

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

Asymptotic significance

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

$$
\begin{aligned}
Z &= \sqrt{q_0} \\
&= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}
\end{aligned}
$$

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

Essentially same as in:

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

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

Asimov approximation for median significance

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

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

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

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

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

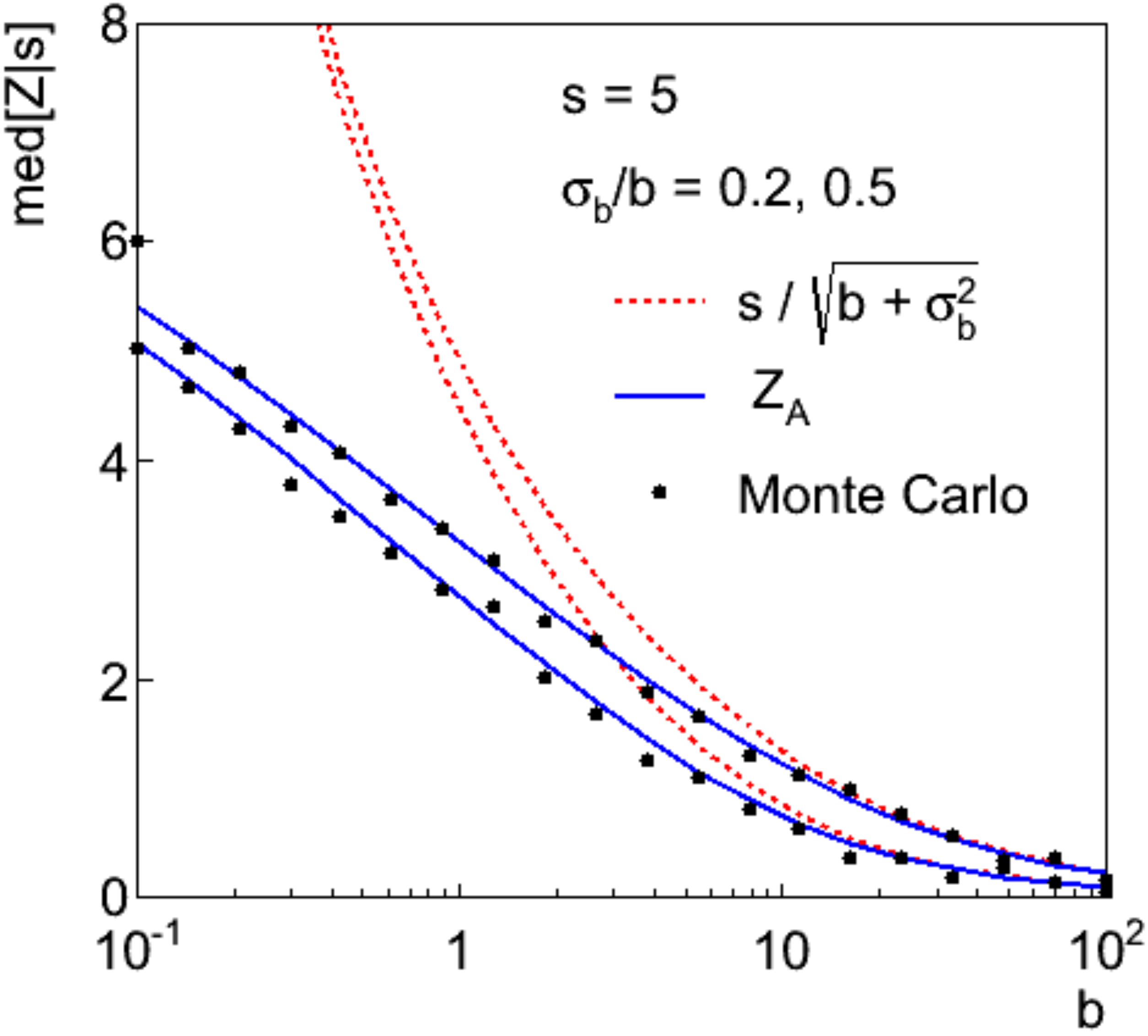
Limiting cases

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

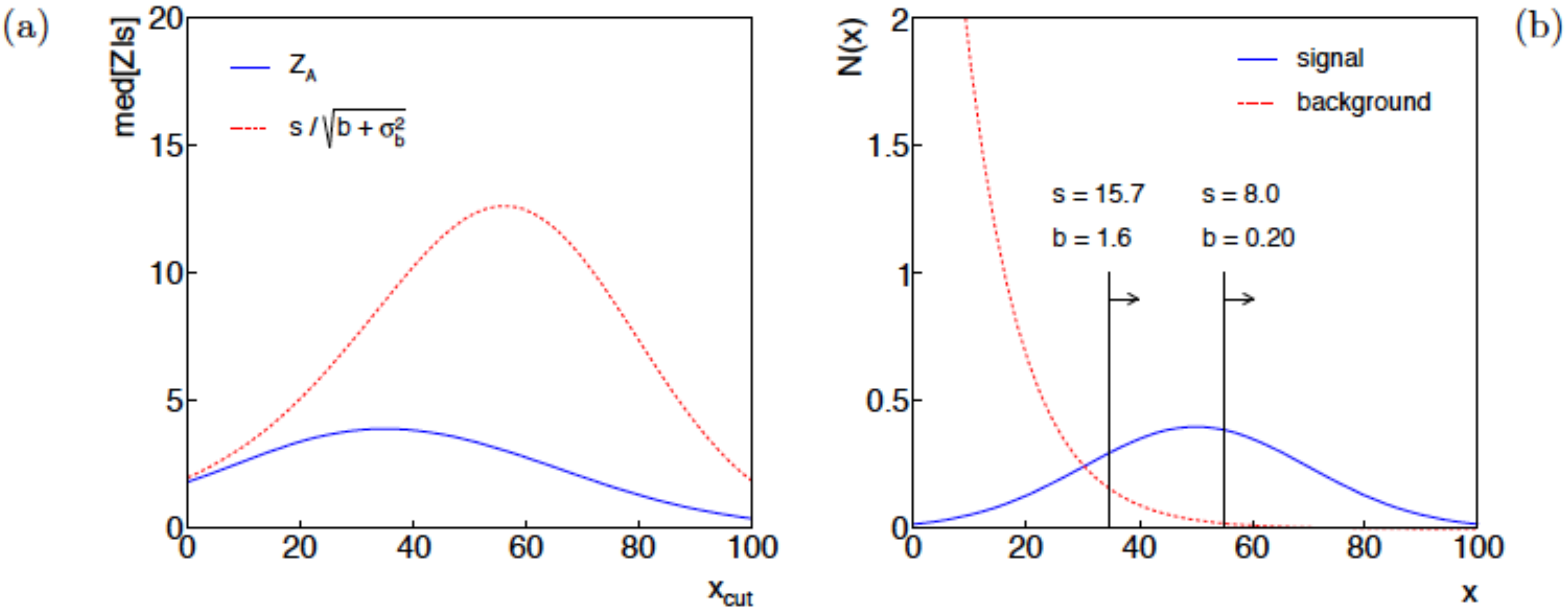


Figure 1: (a) The expected significance as a function of the cut value x_{cut}; (b) the distributions of signal and background with the optimal cut value indicated.

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).

Summary and conclusions

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

HEP has focused on frequentist tests for both *p*-values and limits.

We are very concerned with reporting *p*-values accurately, so worry a lot about systematic uncertainties, nuisance parameters.

Often rely heavily on asymptotic distributions for tests.

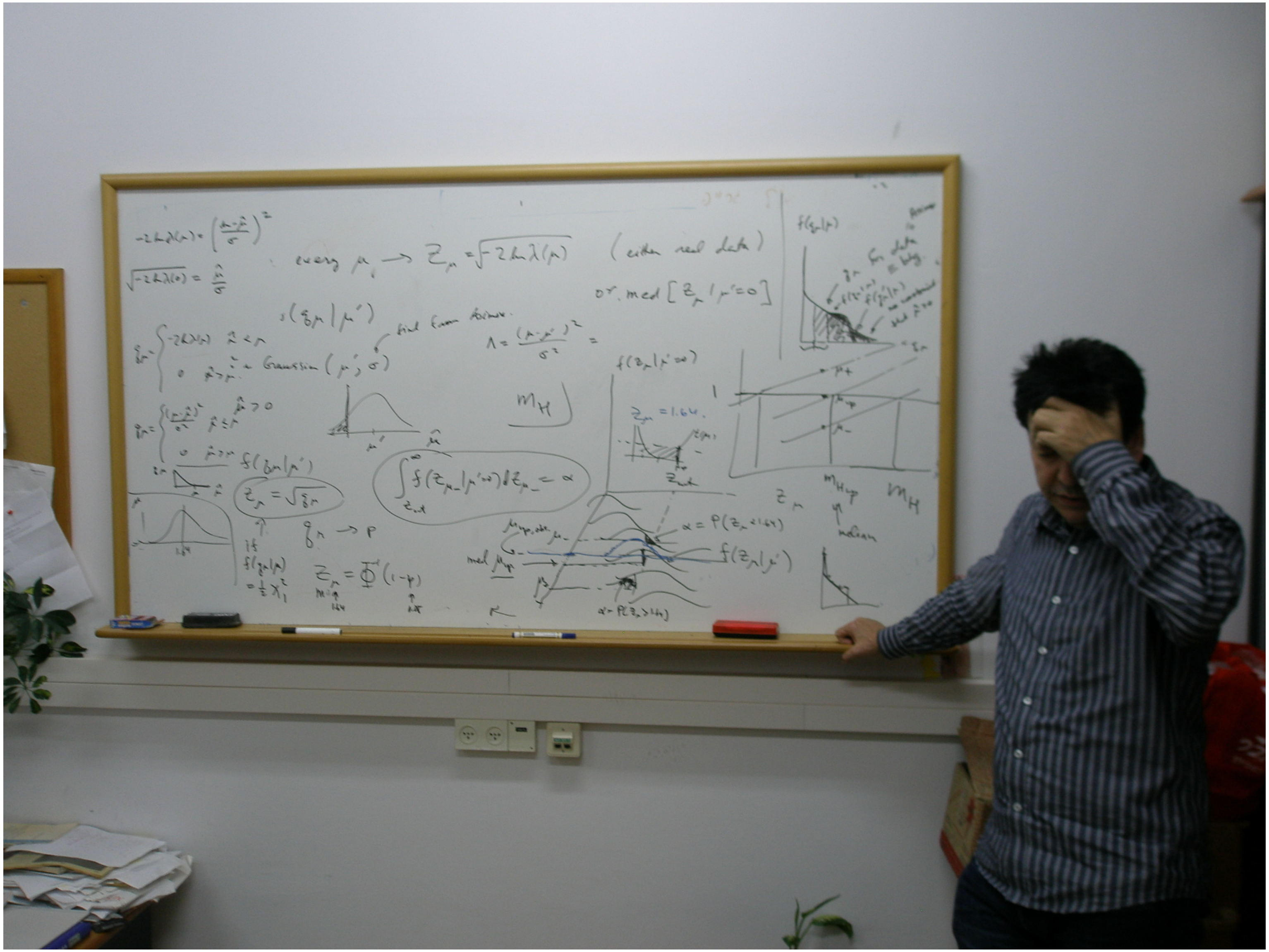
Many important questions untouched in this talk, e.g.,

Corrections for Look-Elsewhere Effect (EG, OV),

Use of Bayesian methods for both limits and discovery,

Unfolding (deconvolution),...

תודה Eilam!



every μ → Z_μ = √(−2 ln λ(μ))
(either real data)
or med [Z_μ | μ' = 0]
find from formula
Λ = (μ−μ')²/σ²
f(Z_μ | μ' = 0)
Gaussian (μ', σ)
m_H
Z_μ = 1.64
Z_μ = √q_μ
q_μ → p
∫ f(Z_μ− | μ' = 0) dZ_μ− = α
Z_μ = Φ⁻¹(1 − p)
α = P(Z_μ > 1.64)
f(Z_μ | μ')
f(q_μ | μ)
m_H,up
median
m_H

CLs
exclusion
(R sinθ cosφ, R sinθ sinφ, R cosθ)
Poiss(n|λ)
$e^{-\lambda}\lambda^n$
q_μ
$\hat{\mu}$
$f(q_\mu|\mu'=0)$
q_-
q_{med}
q_+
$f(q_\mu|\mu'=0)$
$f(q_\mu|\mu)$
$f(q_\mu|\mu) =$
$Z_\mu = \sqrt{q_\mu}$
q_0^{Asimov}
$f(q_\mu|\mu)$

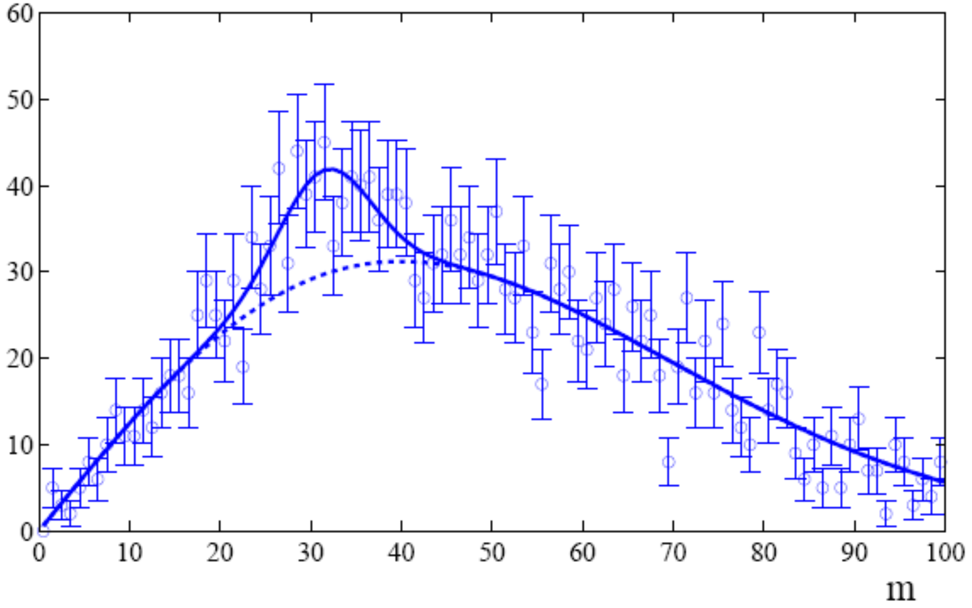
Extra slides

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

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

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

The resulting *p*-value

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

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

Global *p*-value

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

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

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

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

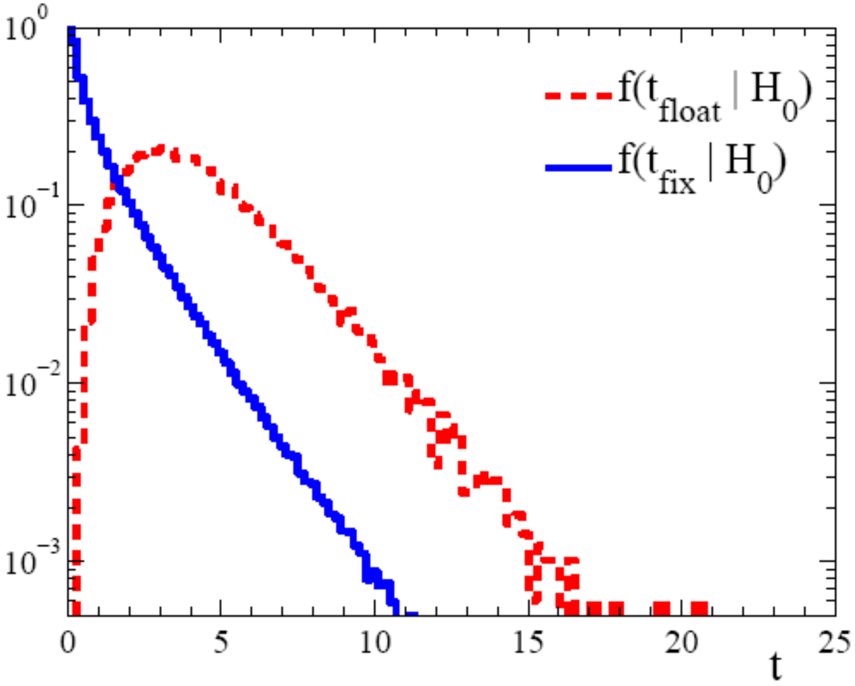
(Note *m* does not appear in the $\mu = 0$ model.)

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

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).

Gross and Vitells

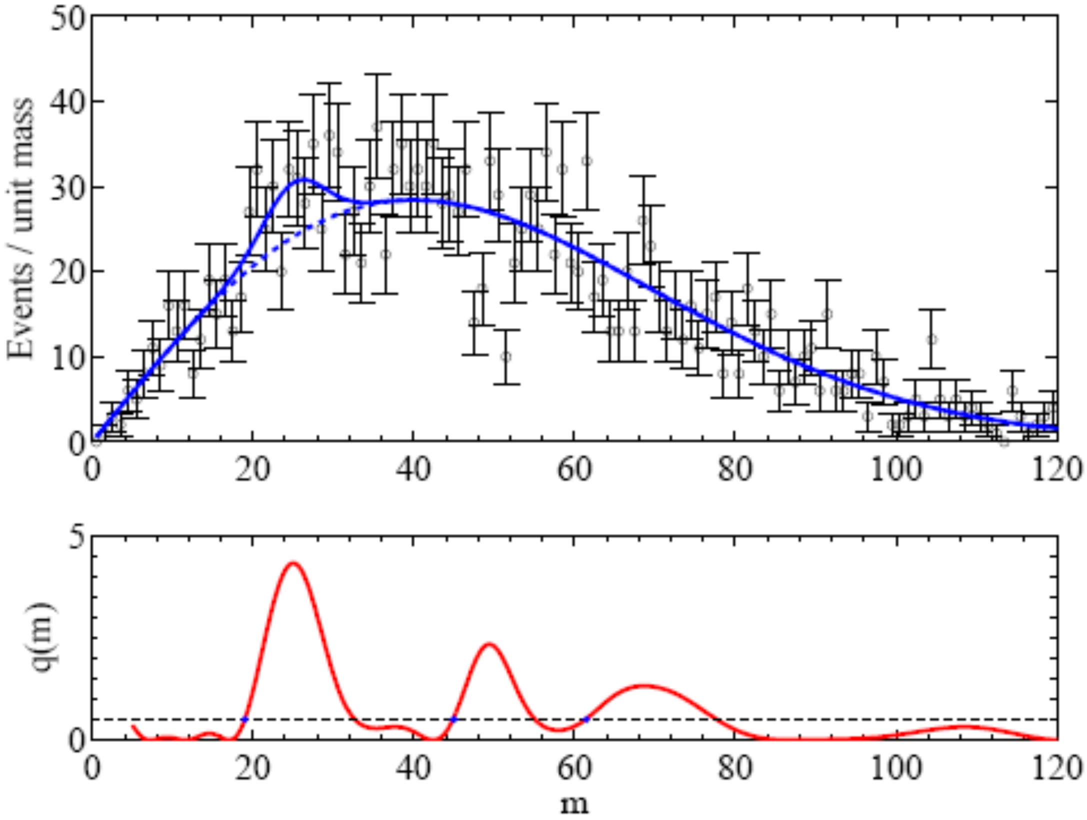
Upcrossings of −2ln$L$

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

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

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

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



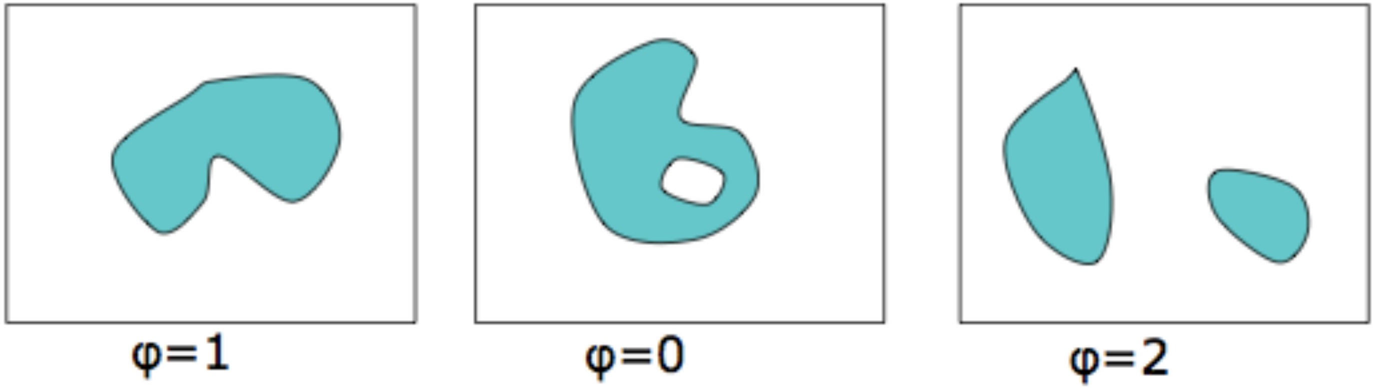
Vitells and Gross, Astropart. Phys. 35 (2011) 230-234; arXiv:1105.4355

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.

Toy study on variable selection

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).

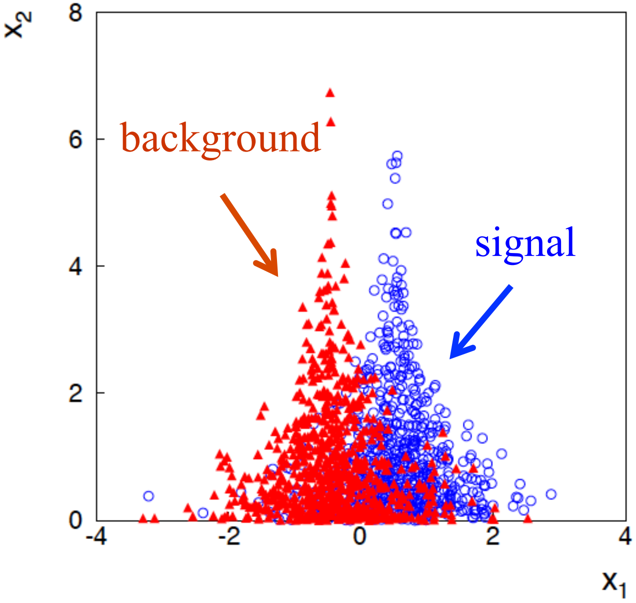
$f(x_1|x_2)$ ~ Gaussian, different means for s/b,
Gaussians have same σ, which depends on x_2,
$f(x_2)$ ~ exponential, same for both s and b,
$f(x_1, x_2) = f(x_1|x_2)\, f(x_2)$:

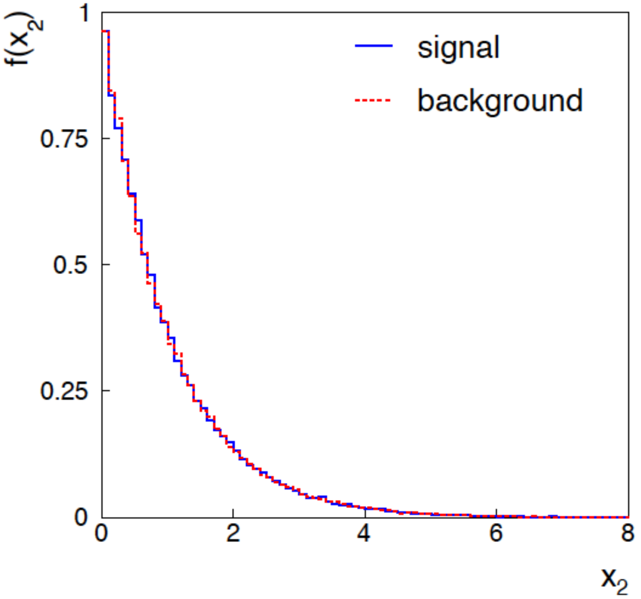
$$f(x_1, x_2|\mathrm{s}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{s})^2/2\sigma^2(x_2)} \, \frac{1}{\lambda} e^{-x_2/\lambda}$$

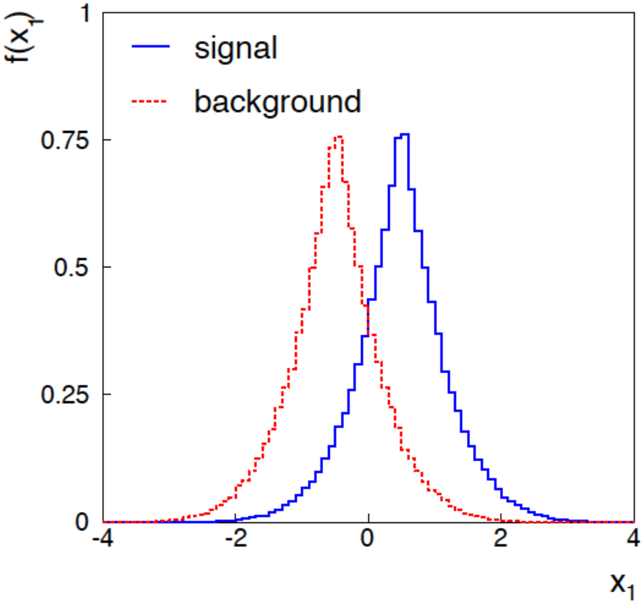
$$f(x_1, x_2|\mathrm{b}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1-\mu_\mathrm{b})^2/2\sigma^2(x_2)} \, \frac{1}{\lambda} e^{-x_2/\lambda}$$

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

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:

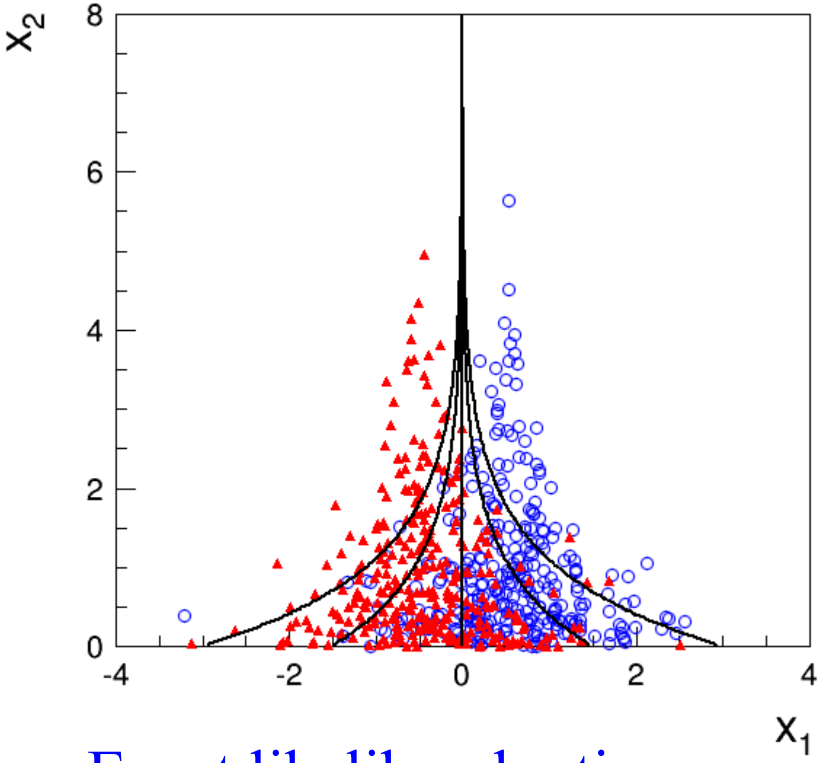
$$t(x_1, x_2) = \frac{f(x_1, x_2|\mathrm{s})}{f(x_1, x_2|\mathrm{b})}$$

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

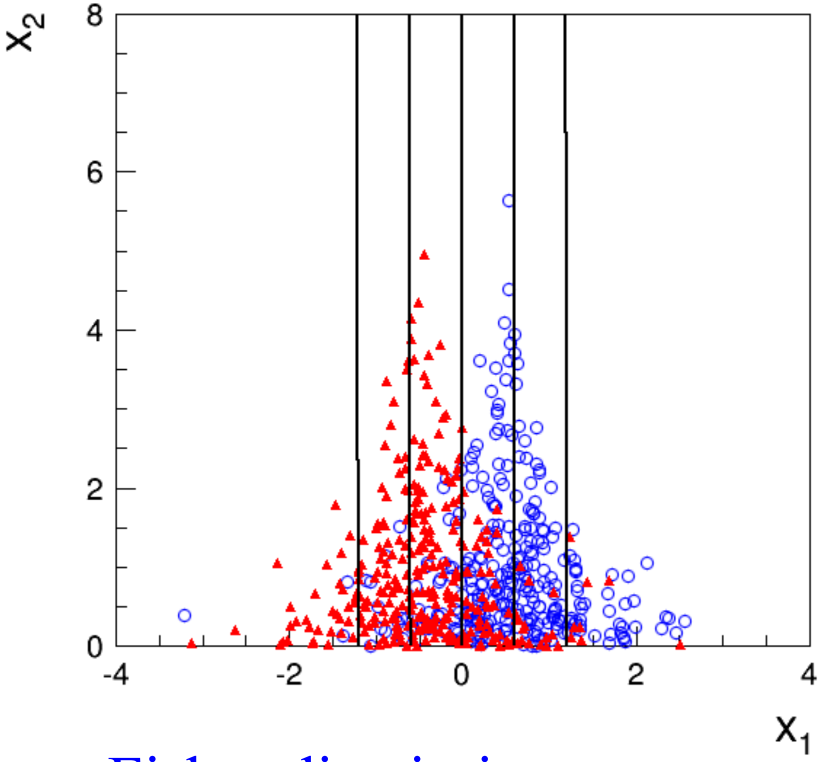
$$\ln t = \frac{\frac{1}{2}(\mu_\mathrm{b}^2 - \mu_\mathrm{s}^2) + (\mu_\mathrm{s} - \mu_\mathrm{b})x_1}{\sigma_0^2 e^{-2x_2/\xi}}$$

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

Contours of constant classifier output

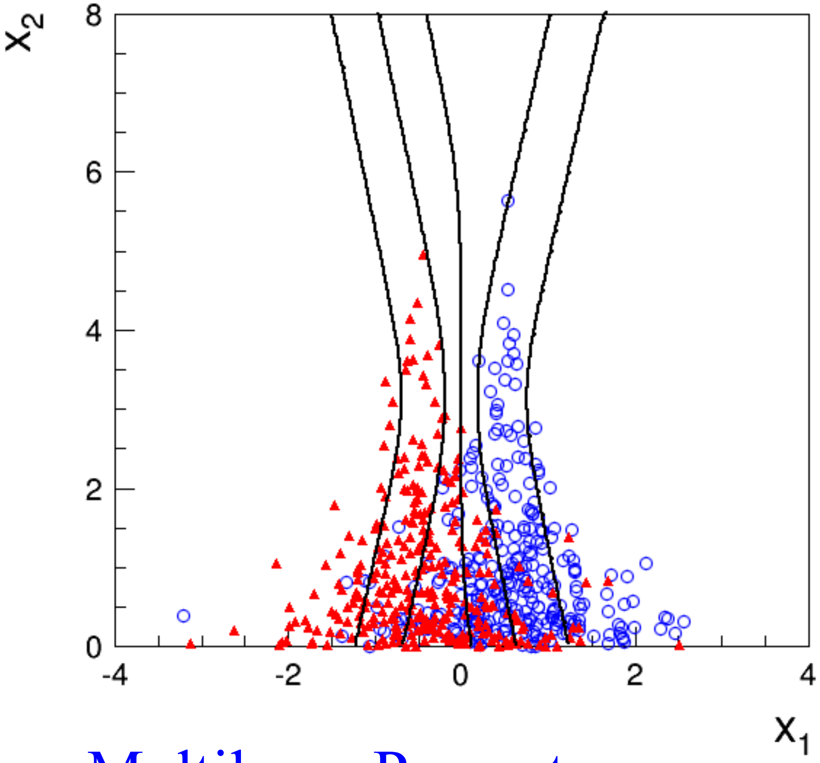


Exact likelihood ratio

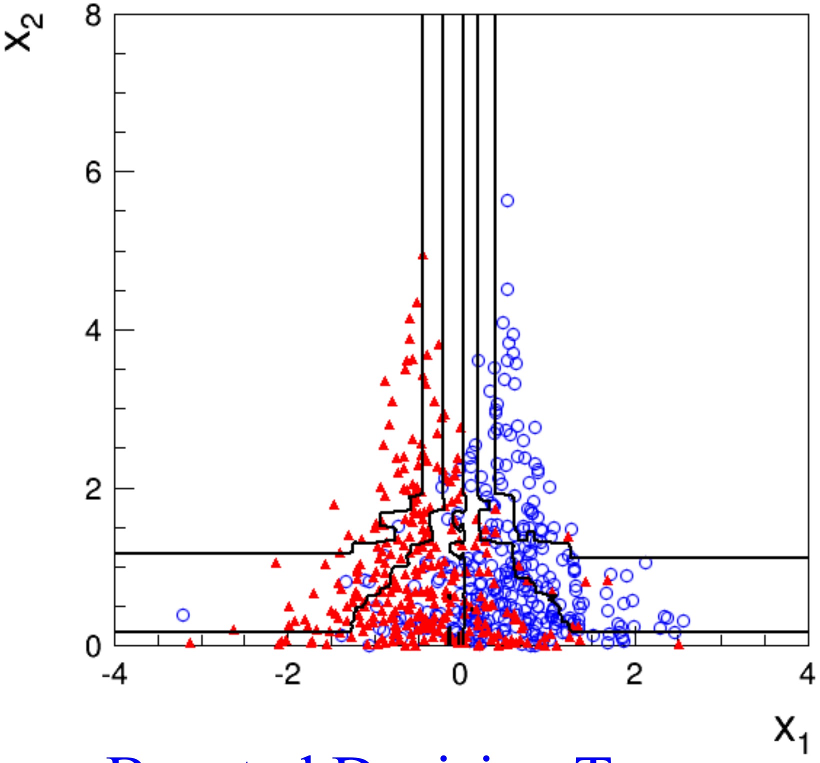


Fisher discriminant

Contours of constant classifier output



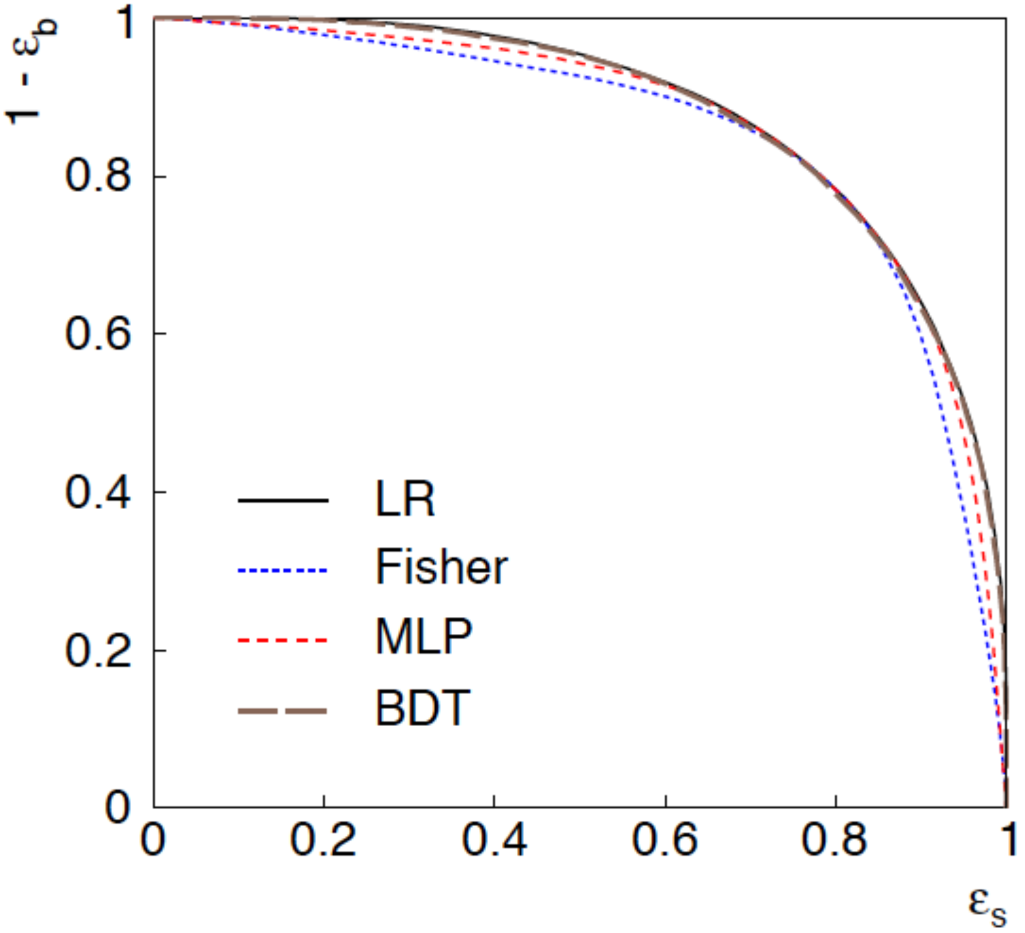
Multilayer Perceptron
1 hidden layer with 2 nodes



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.