

Recent developments in statistical methods for particle physics



Particle Physics Seminar
Warwick, 17 February 2011



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

Outline

Large-sample statistical formulae for a search at the LHC

Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727,
EPJC 71 (2011) 1-19

Significance test using profile likelihood ratio

Systematics included via nuisance parameters

Distributions in large sample limit, no MC used.

Progress on related issues (some updates from PHYSTAT2011):

The “look elsewhere effect”

The “CLs” problem

Combining measurements

Improving treatment of systematics

Prototype search analysis

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

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

Assume the n_i are Poisson distributed with expectation values

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

strength parameter

where

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

signal

background

Prototype analysis (II)

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

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

Assume the m_i are Poisson distributed with expectation values

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

 nuisance parameters ($\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}}$)

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:

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

maximizes L for Specified μ

maximize L

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

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

Test statistic for discovery

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

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

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

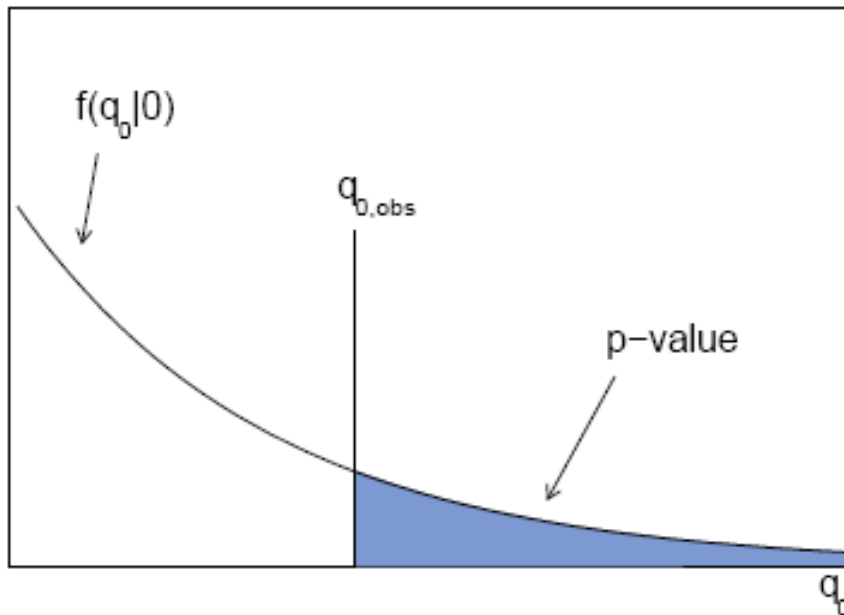
Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

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

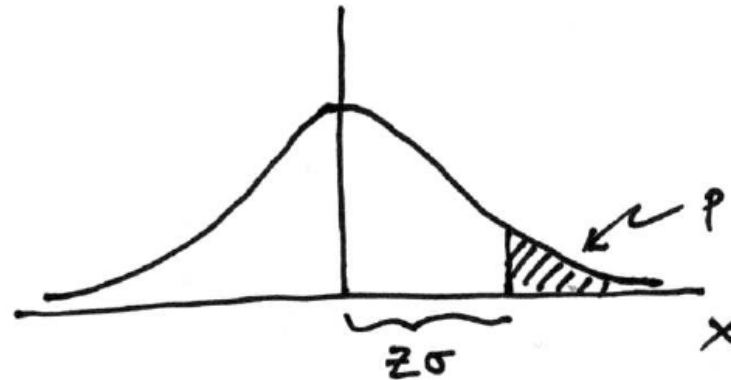


From p -value get equivalent significance,

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

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.

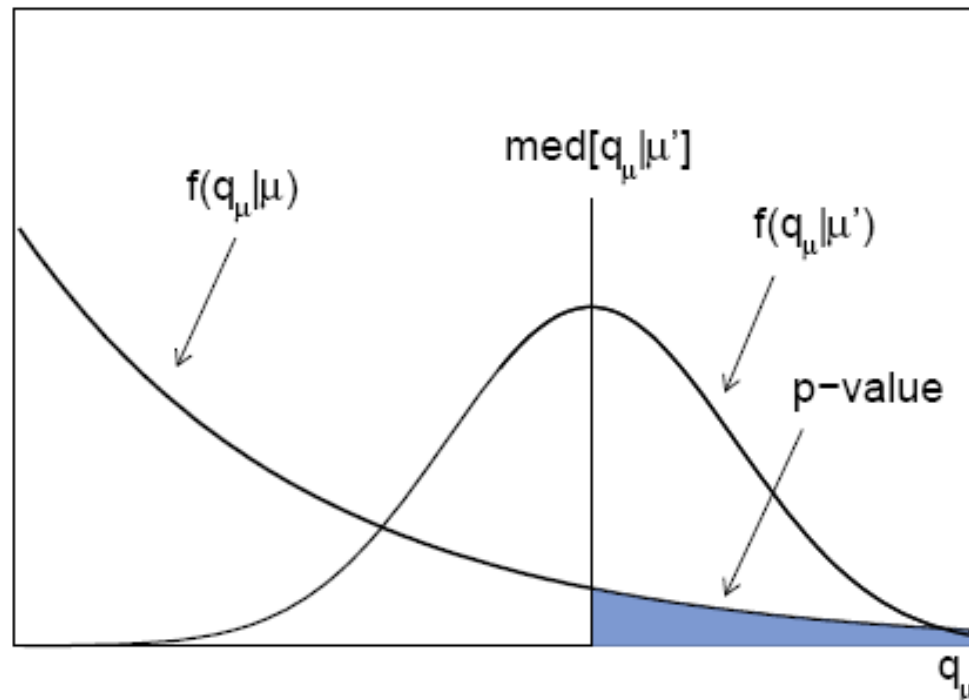


$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \text{1 - TMath::Freq}$$

$$Z = \Phi^{-1}(1 - p) \quad \text{TMath::NormQuantile}$$

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 $f(q_0|\mu')$,

Test statistic for upper limits

For purposes of setting an upper limit on μ use

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

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

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

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

Alternative test statistic for upper limits

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

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

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

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

q_μ is simpler in important ways: asymptotic distribution is independent of nuisance parameters.

Wald approximation for profile likelihood ratio

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

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

Use approximation due to Wald (1943)

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

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

 sample size

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

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

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

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

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

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

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

The Asimov data set

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

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

$$m_i = u_i$$

→ $\hat{\mu} = \mu' \quad \hat{\theta} = \theta$

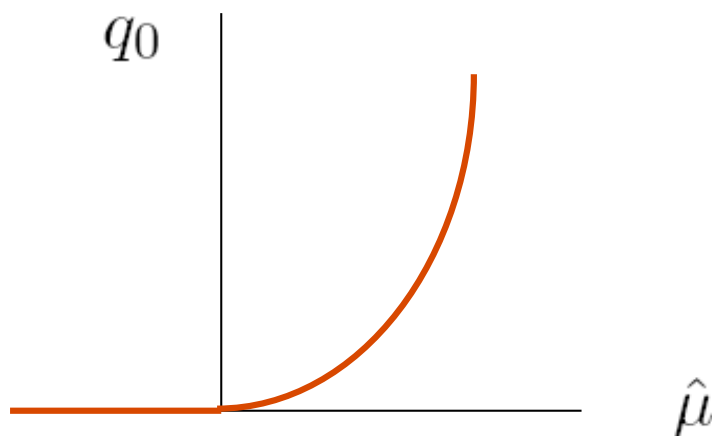
$$\lambda_A(\mu) = \frac{L_A(\mu, \hat{\theta})}{L_A(\hat{\mu}, \hat{\theta})} = \frac{L_A(\mu, \hat{\theta})}{L_A(\mu', \theta)}$$

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

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.,

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

Distribution of q_0

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

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

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

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

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

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

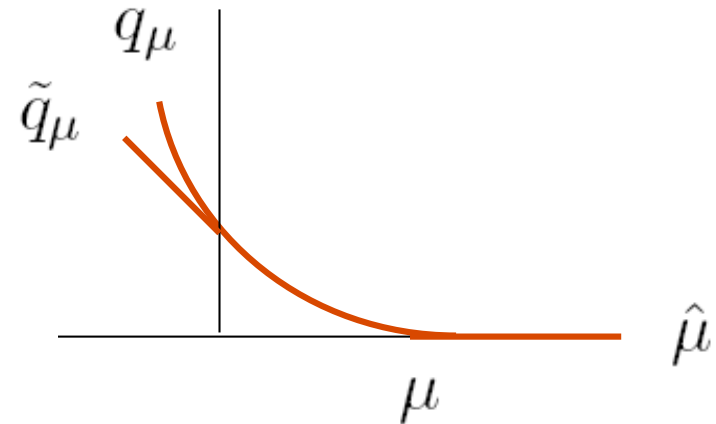
Therefore the discovery significance Z is simply

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

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

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

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



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

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

Distribution of q_μ

Similar results for q_μ

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

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

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

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

Distribution of \tilde{q}_μ

Similar results for \tilde{q}_μ

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

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

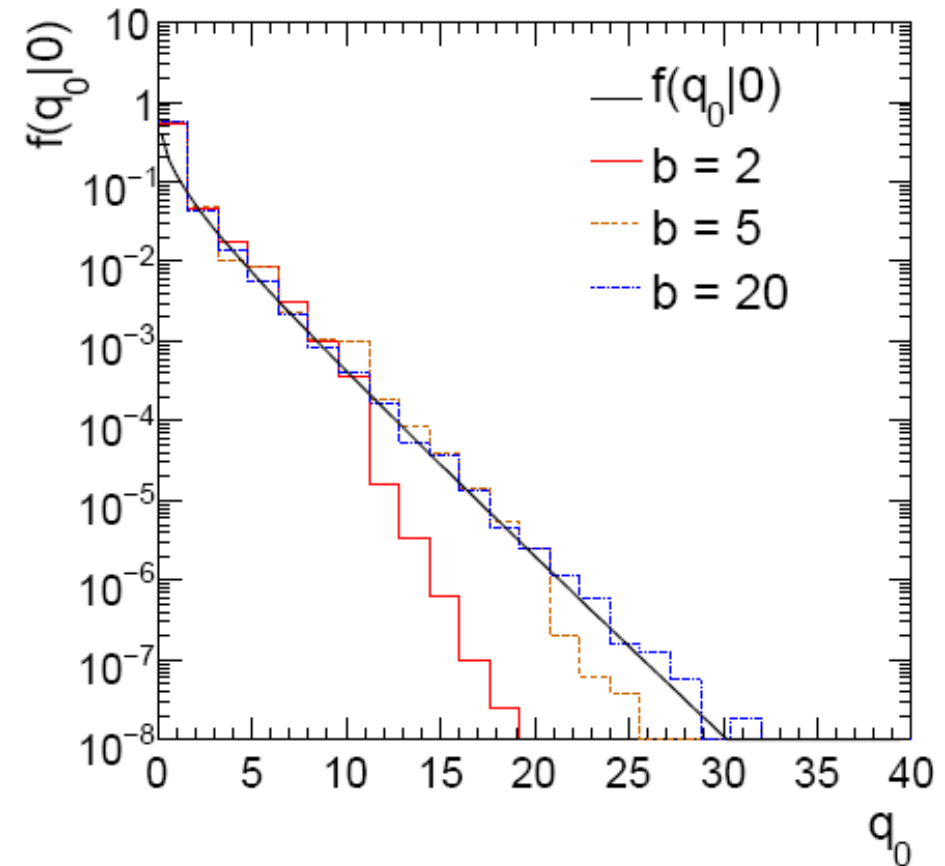
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

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

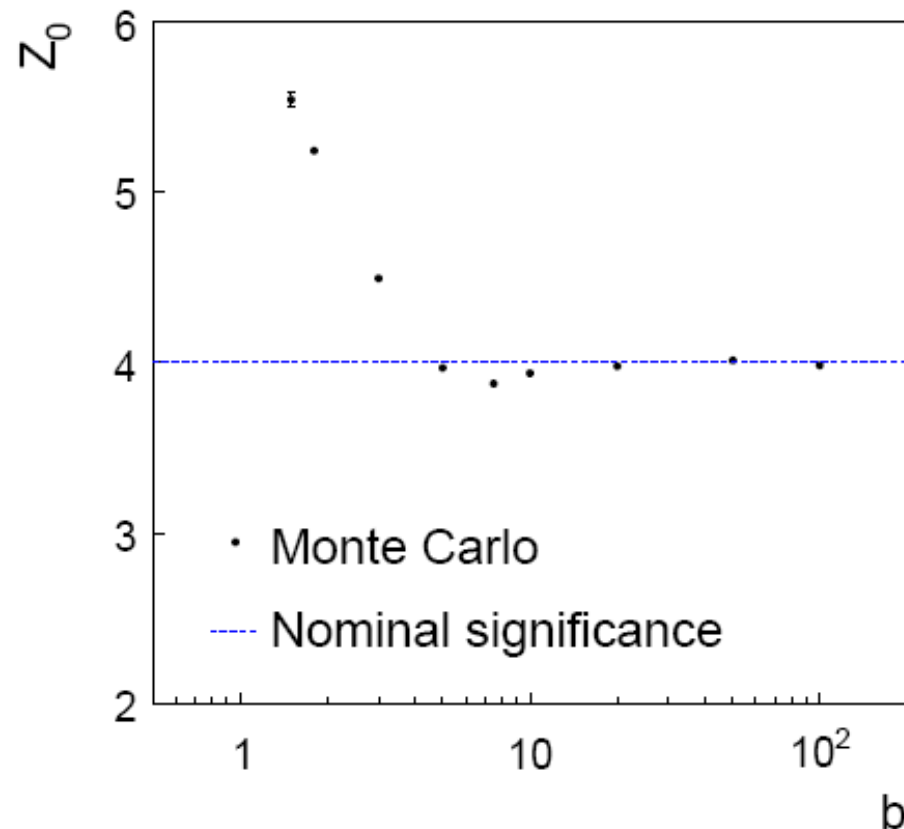


Monte Carlo test of asymptotic formulae

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

For very low b , asymptotic formula underestimates Z_0 .

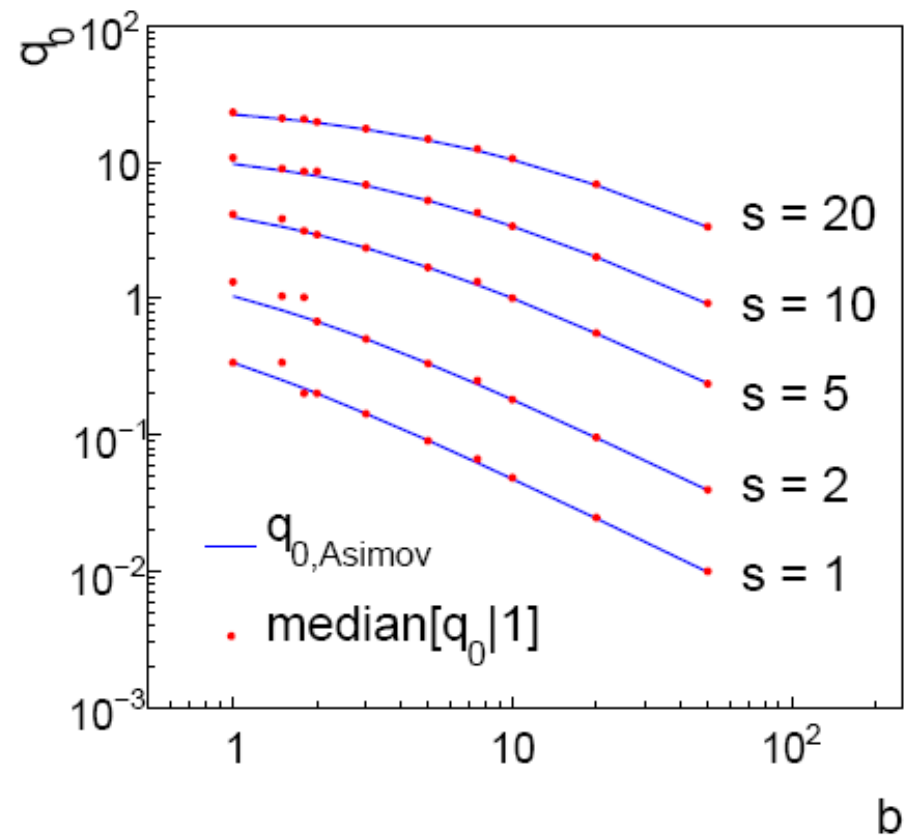
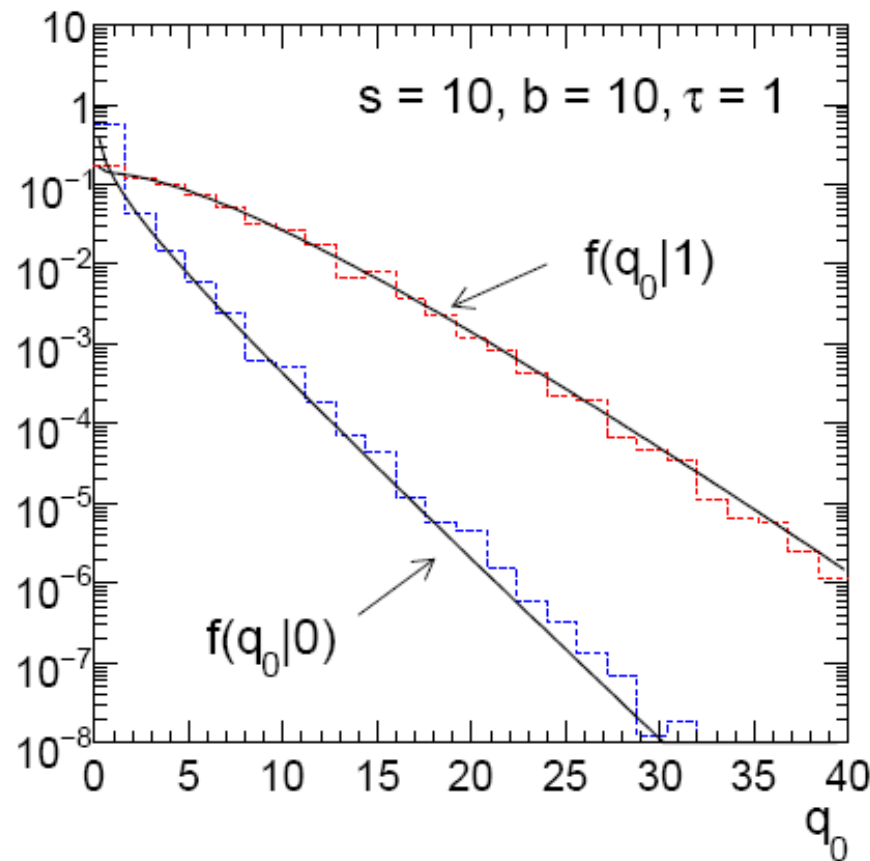
Then slight overshoot before rapidly converging to MC value.



Monte Carlo test of asymptotic formulae

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

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



Monte Carlo test of asymptotic formulae

Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$

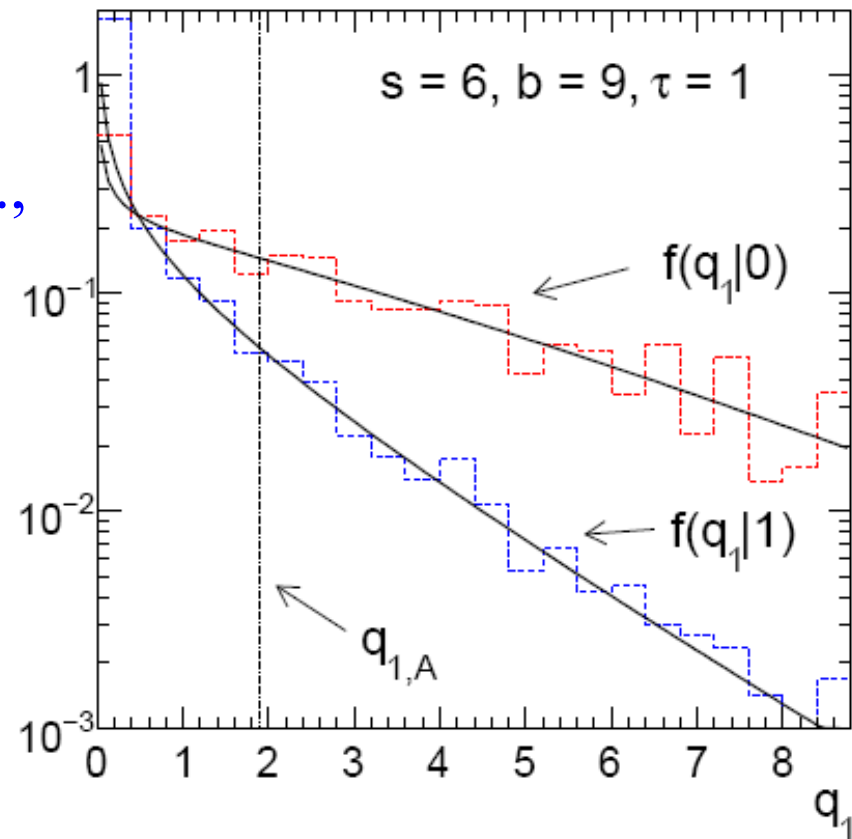
Use q_μ to find p -value of hypothesized μ values.

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

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

Median[$q_1 | 0$] gives “exclusion sensitivity”.

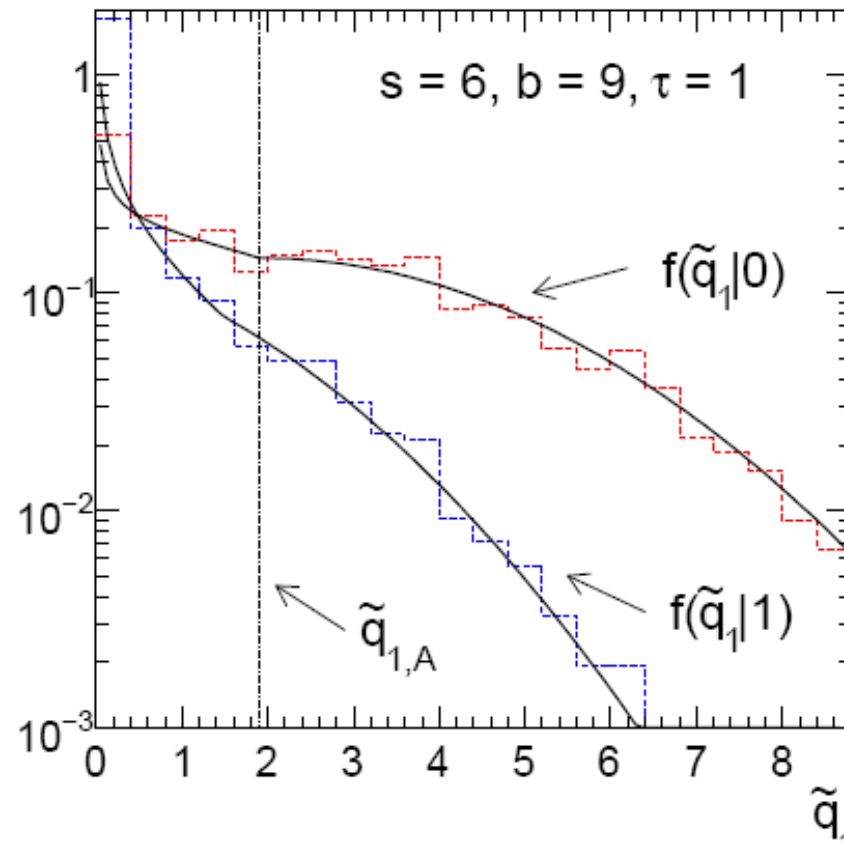
Here asymptotic formulae good
for $s = 6$, $b = 9$.



Monte Carlo test of asymptotic formulae

Same message for test based on \tilde{q}_μ .

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



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

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

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

Significance for rejecting $s = 0$ is therefore

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

Expected (median) significance assuming signal rate s is

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

Better approximation for Poisson significance

Likelihood function for parameter s is

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

or equivalently the log-likelihood is

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

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

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

Approximate Poisson significance (continued)

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

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

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

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

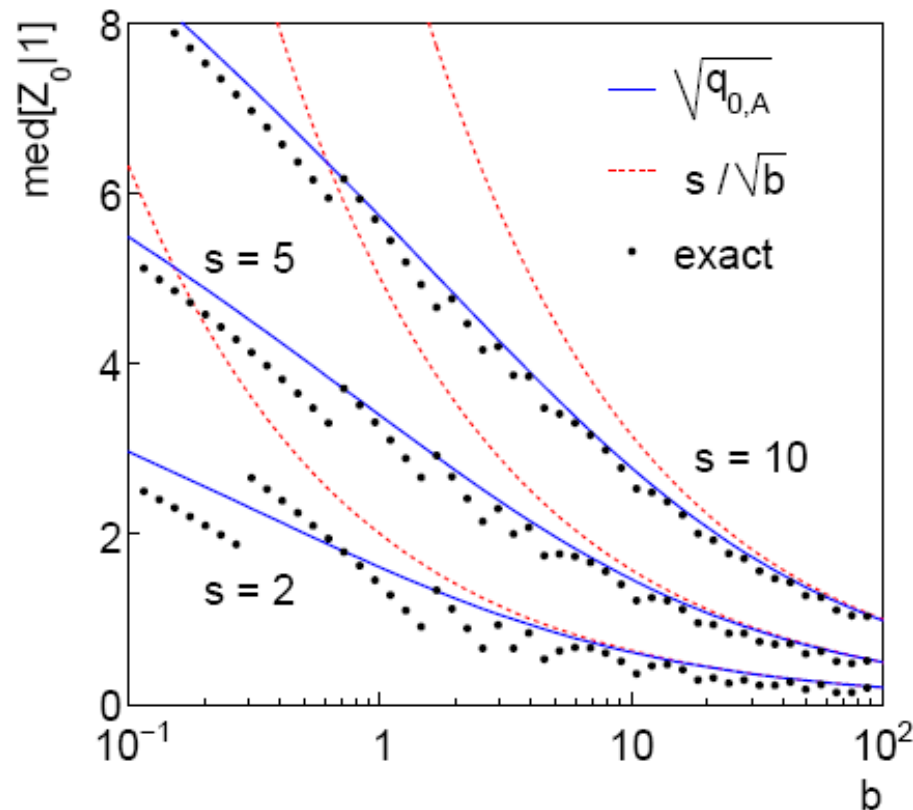
To find $\text{median}[Z_0|s+b]$, let $n \rightarrow s + b$ (i.e., the Asimov data set):

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

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

$n \sim \text{Poisson}(\mu s + b)$, median significance,
assuming $\mu = 1$, of the hypothesis $\mu = 0$

CCGV, arXiv:1007.1727



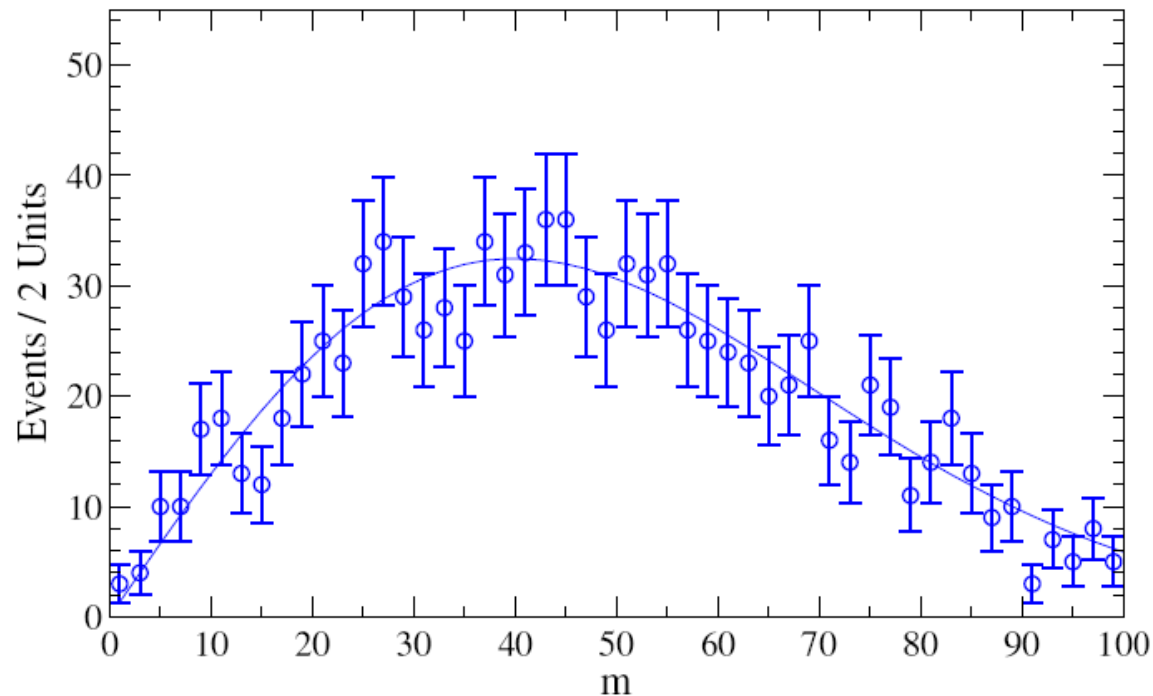
“Exact” values from MC,
jumps due to discrete data.

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

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

Example 2: Shape analysis

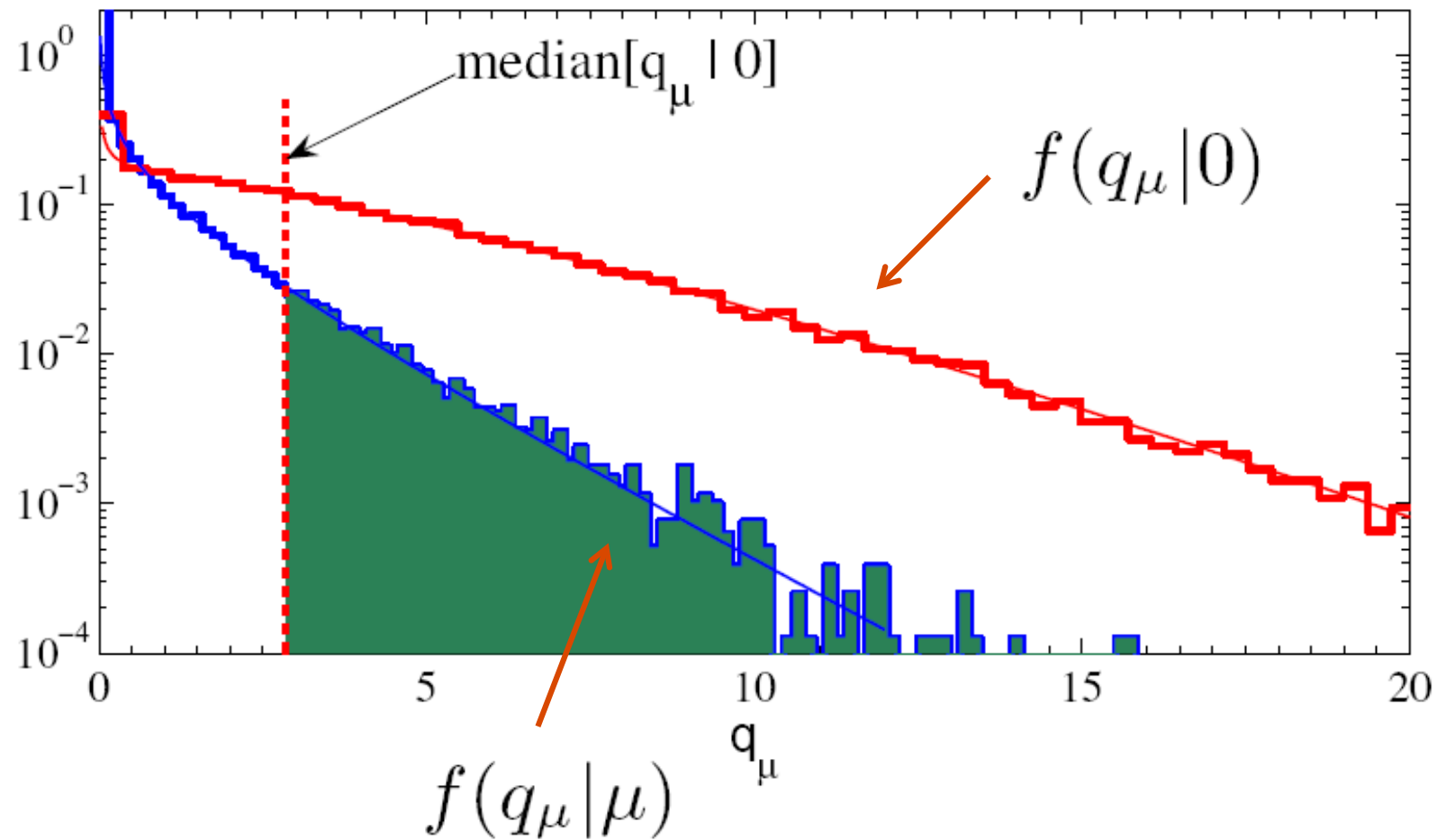
Look for a Gaussian bump sitting on top of:



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

Monte Carlo test of asymptotic formulae

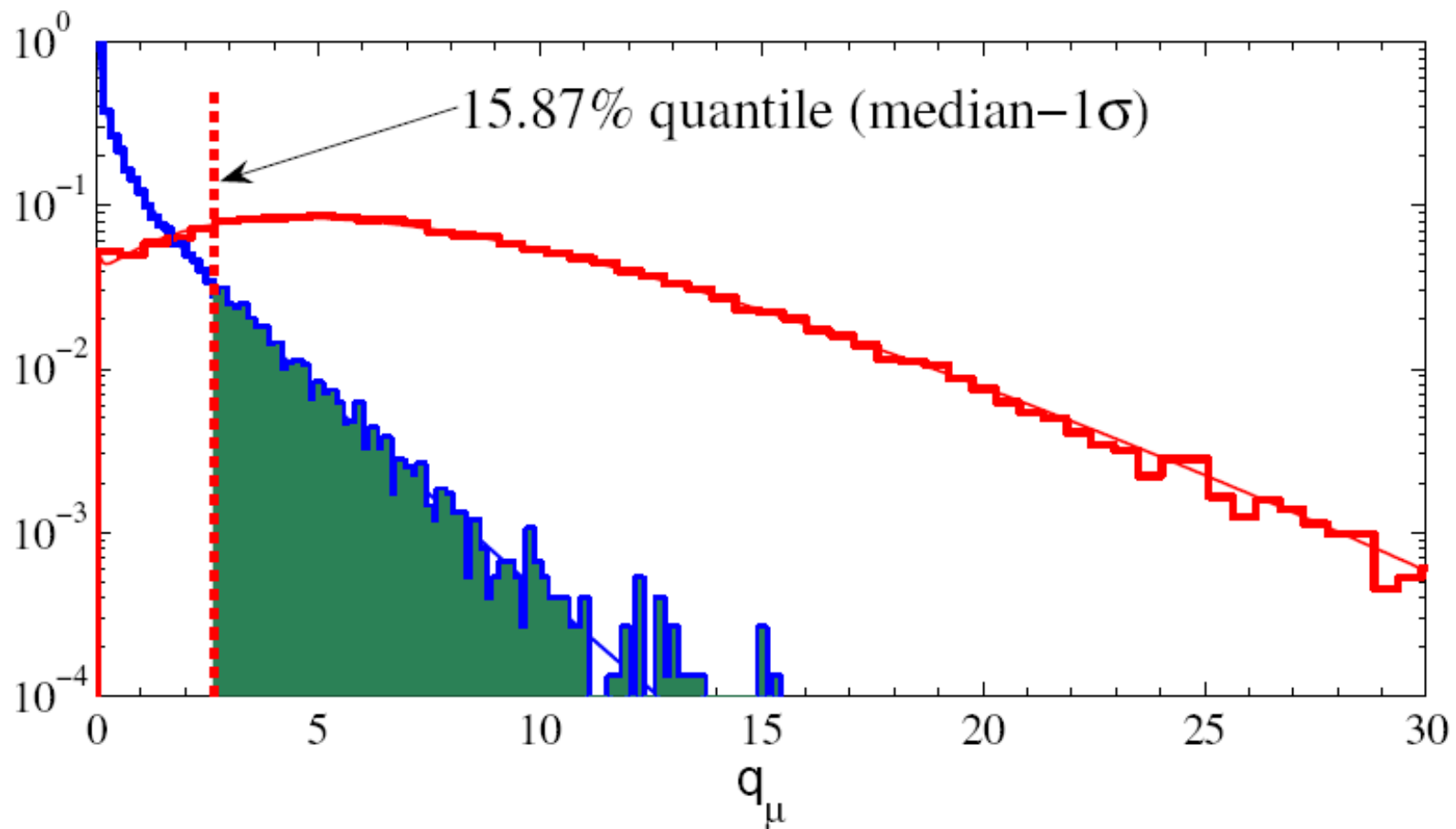
Distributions of q_μ here for μ that gave $p_\mu = 0.05$.



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

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

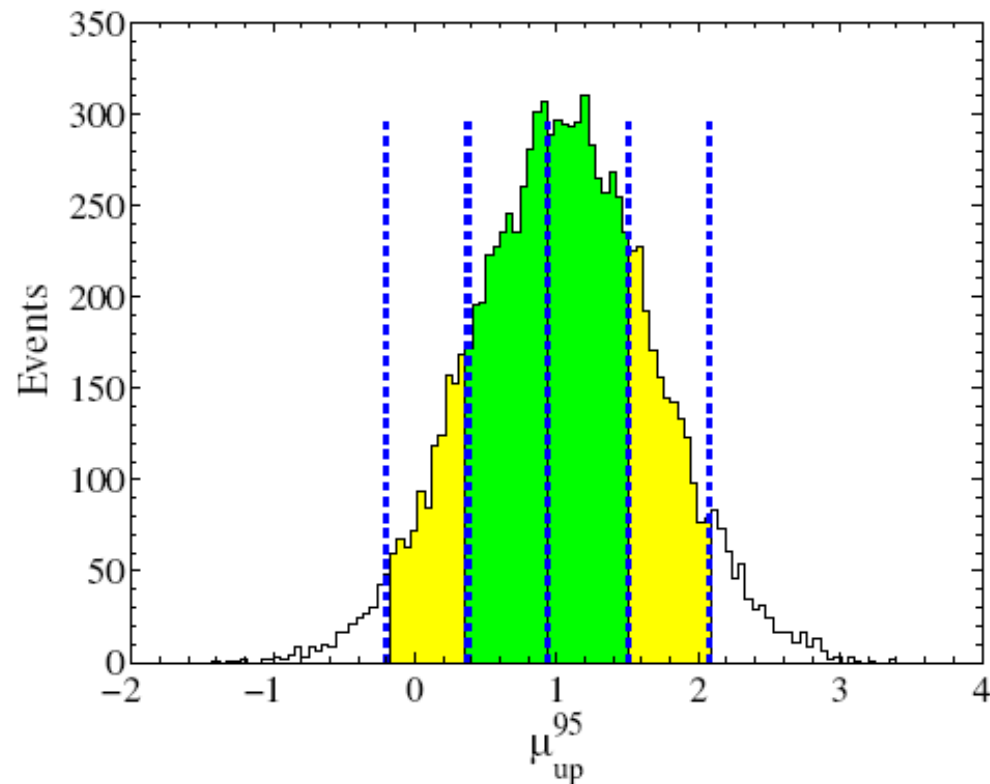
But we have full $f(q_\mu|0)$; we can get any desired quantiles.



Distribution of upper limit on μ

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

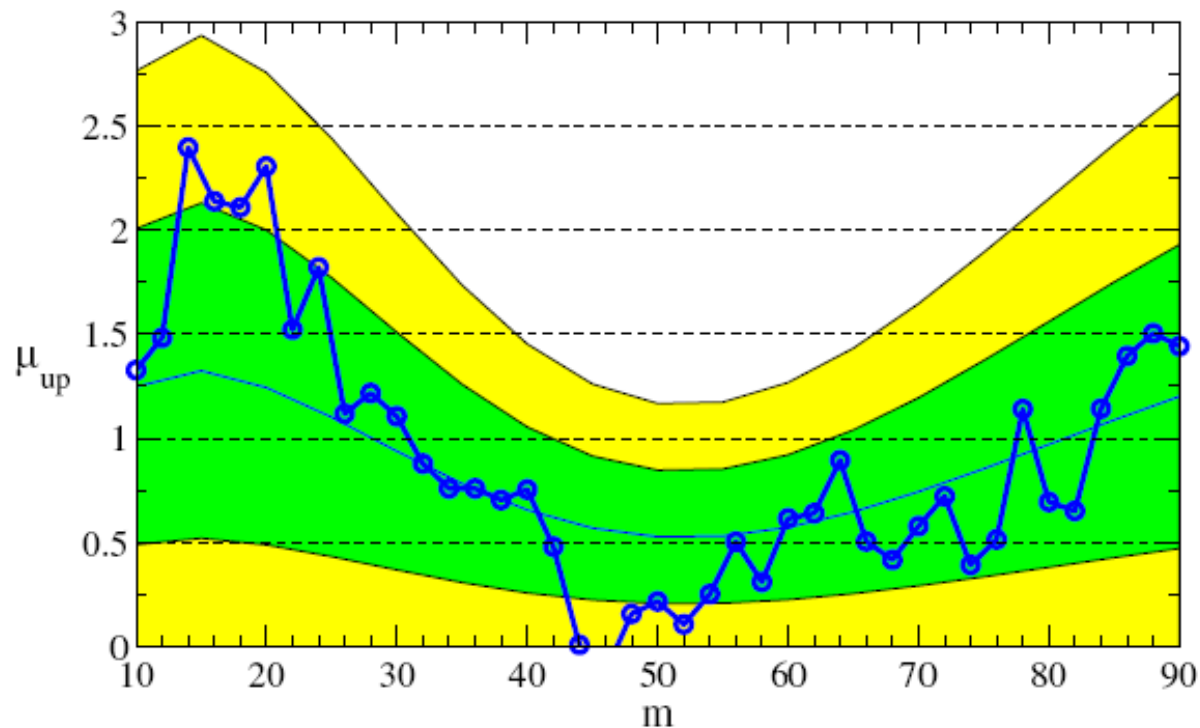
Vertical lines from asymptotic formulae



Limit on μ versus peak position (mass)

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


Points are from a single arbitrary data set.




Using likelihood ratio L_{s+b}/L_b

Many searches at the Tevatron have used the statistic

$$q = -2 \ln \frac{L_{s+b}}{L_b}$$

likelihood of $\mu = 1$ model (s+b) 

likelihood of $\mu = 0$ model (bkg only) 

This can be written

$$q = -2 \ln \frac{L(\mu = 1, \hat{\boldsymbol{\theta}}(1))}{L(\mu = 0, \hat{\boldsymbol{\theta}}(0))} = -2 \ln \lambda(1) + 2 \ln \lambda(0)$$

Wald approximation for L_{s+b}/L_b

Assuming the Wald approximation, q can be written as

$$q = \frac{(\hat{\mu} - 1)^2}{\sigma^2} - \frac{\hat{\mu}^2}{\sigma^2} = \frac{1 - 2\hat{\mu}}{\sigma^2}$$

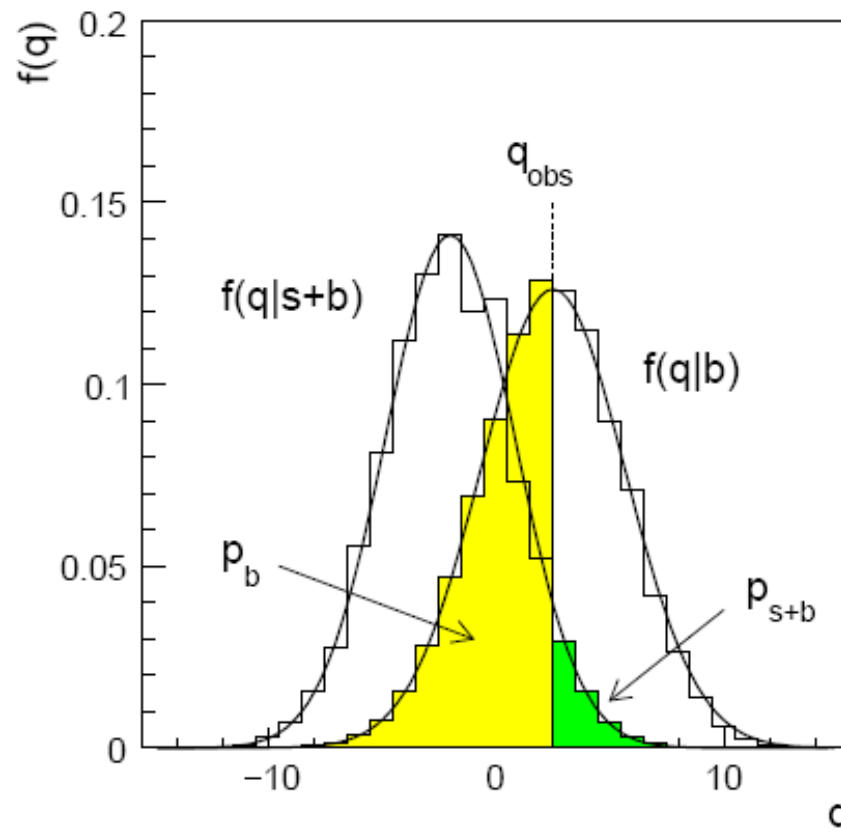
i.e. q is Gaussian distributed with mean and variance of

$$E[q] = \frac{1 - 2\mu}{\sigma^2} \quad V[q] = \frac{4}{\sigma^2}$$

To get σ^2 use 2nd derivatives of $\ln L$ with Asimov data set.

Example with L_{s+b}/L_b

Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$
 $b = 20$, $s = 10$, $\tau = 1$.



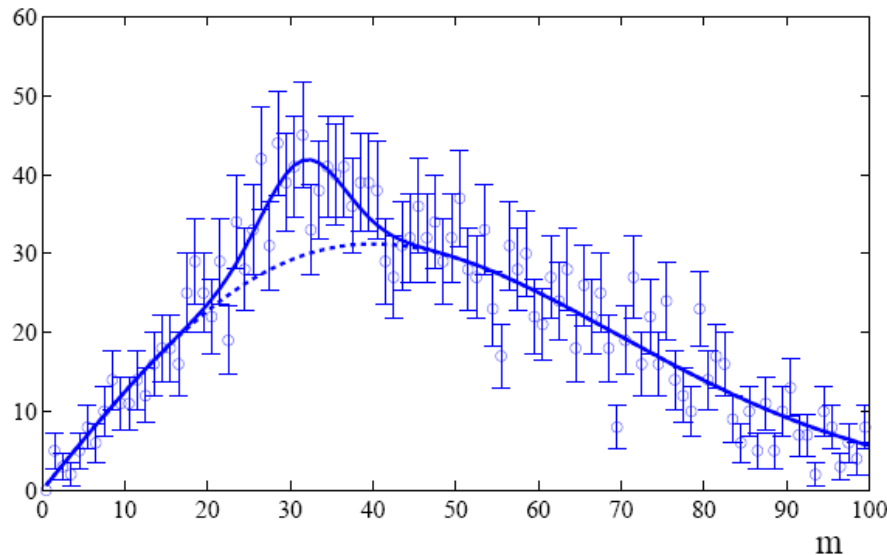
So even for smallish data sample, Wald approximation can be useful; no MC needed.

The Look-Elsewhere Effect

Eilam Gross and Ofer Vitells, arXiv:1005.1891 (\rightarrow EPJC)

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?

p -value for fixed mass

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

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

gives the probability to find a value of t_{fix} at least as great as observed at the specific mass m_0 .

p-value for floating mass

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

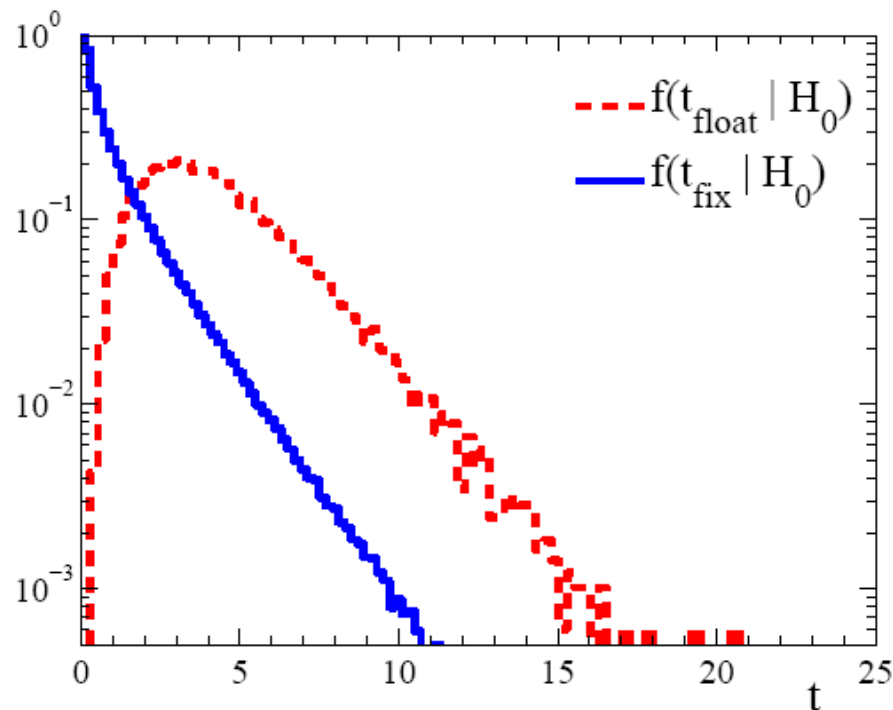
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \quad (\text{Note } m \text{ does not appear in the } \mu = 0 \text{ model.})$$

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

Distributions of t_{fix} , t_{float}

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

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



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

So getting t_{float} distribution is more difficult.

Trials factor

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

Gross and Vitells (arXiv:1005.1891) show that the “trials factor” can be approximated by

$$F_{\text{trials}} \equiv \frac{p_{\text{float}}}{p_{\text{fix}}} \approx 1 + \sqrt{\frac{\pi}{2}} \langle \mathcal{N} \rangle Z_{\text{fix}}$$

where $\langle \mathcal{N} \rangle$ = average number of “upcrossings” of $-2\ln L$ in fit range and

$$Z_{\text{fix}} = \Phi^{-1}(1 - p_{\text{fix}}) = \sqrt{t_{\text{fix}}}$$

is the significance for the fixed mass case.

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).

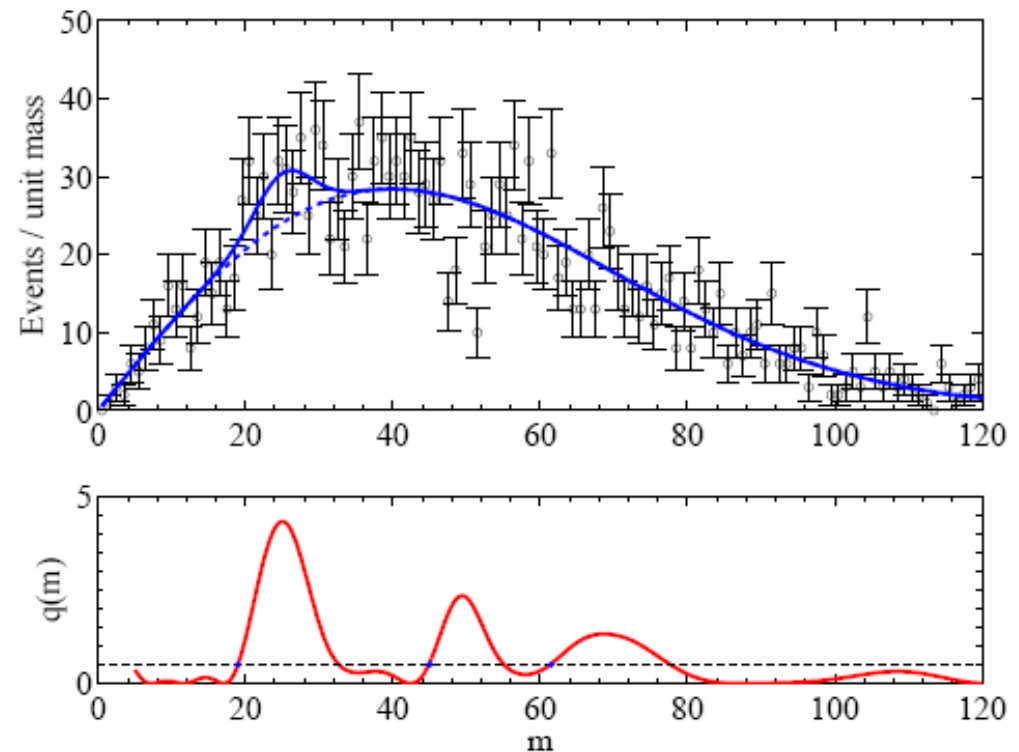
Upcrossings of $-2\ln L$

The Gross-Vitells formula for the trials factor requires the mean number “upcrossings” of $-2\ln L$ in the fit range based on fixed threshold.

$$\begin{aligned} P(q_0 > u) \\ &\leq E[N_u] + P(q_0(0) > u) \\ &= \mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u) \end{aligned}$$



estimate with MC
at low reference
level

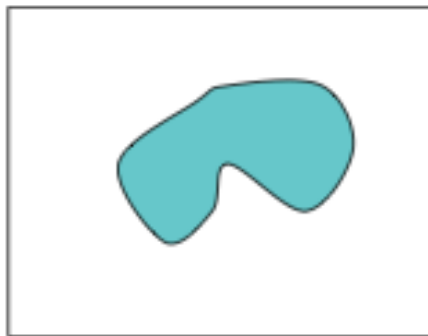


Multidimensional look-elsewhere effect

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

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

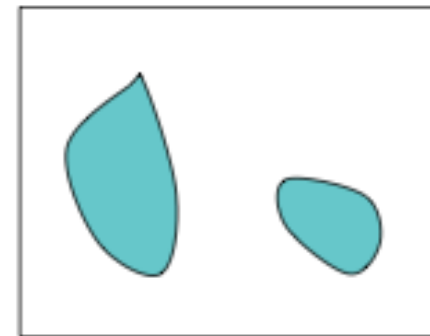
- Number of disconnected components minus number of 'holes'



$\varphi=1$



$\varphi=0$

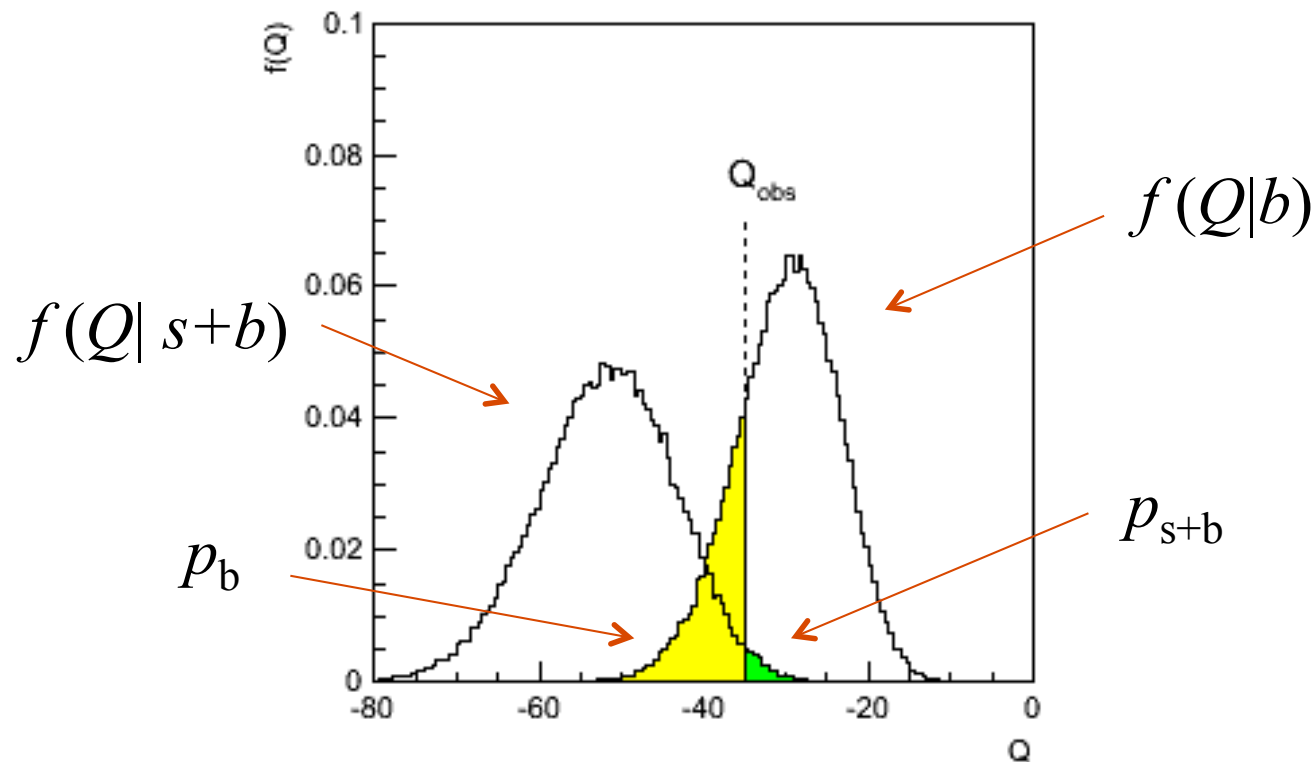


$\varphi=2$

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

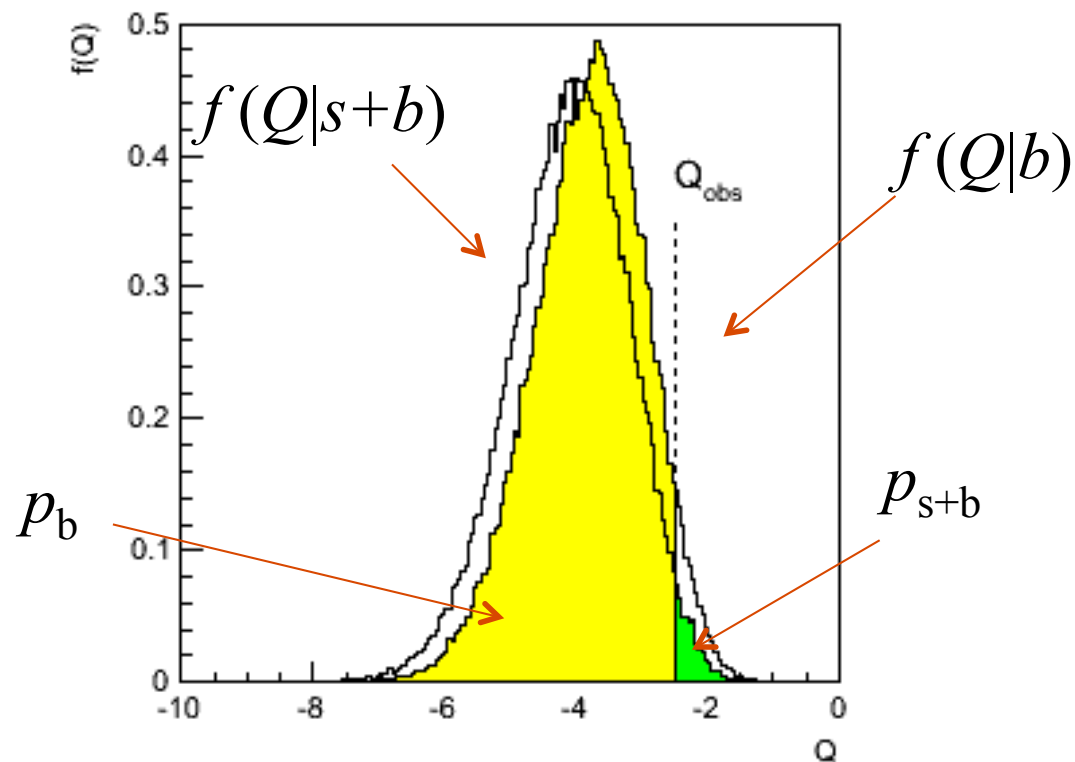
The “CL_s” issue

When the b and $s+b$ hypotheses are well separated, there is a high probability of excluding the $s+b$ hypothesis ($p_{s+b} < \alpha$) if in fact the data contain background only (power of test of $s+b$ relative to the alternative b is high).



The “CL_s” issue (2)

But if the two distributions are close to each other (e.g., we test a Higgs mass far above the accessible kinematic limit) then there is a non-negligible probability of rejecting s+b even though we have low sensitivity (test of s+b low power relative to b).



In limiting case of no sensitivity, the distributions coincide and the probability of exclusion = α (e.g. 0.05).

But we should not regard a model as excluded if we have no sensitivity to it!

The CL_s solution

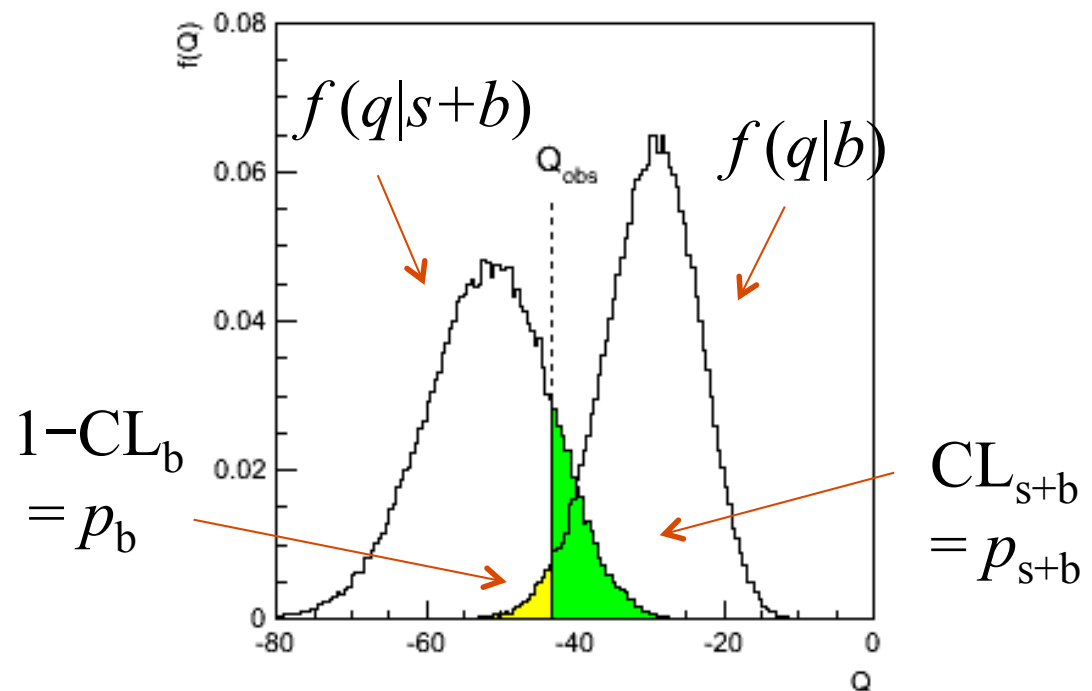
The CL_s solution (A. Read et al.) is to base the test not on the usual p -value (CL_{s+b}), but rather to divide this by CL_b (one minus the background of the b-only hypothesis, i.e.,

Define:

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

Reject $s+b$
hypothesis if:

$$CL_s \leq \alpha$$



Reduces “effective” p -value when the two distributions become close (prevents exclusion if sensitivity is low).

CL_s discussion

In the CLs method the p-value is reduced according to the recipe

$$p_{\mu} \rightarrow \frac{p_{\mu}}{1 - p_b}$$

Statistics community does not smile upon ratio of p-values
An alternative would to regard parameter μ as excluded if:

- (a) p -value of $\mu < 0.05$
- (b) power of test of μ with respect to background-only exceeds a specified threshold

i.e. “**Power Constrained Limits**”. Coverage is $1-\alpha$ if one is sensitive to the tested parameter (sufficient power) otherwise never exclude (coverage is then 100%).

Ongoing study. In any case should produce CL_s result for purposes of comparison with other experiments.

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 $\mu, \boldsymbol{\theta}$.

→ ongoing ATLAS/CMS effort with **RooStats** framework

<https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome>

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

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

RooStats

G. Schott
PHYSTAT2011

a collaborative project with contributors from ATLAS, CMS and ROOT aimed to **provide & consolidate statistical tools** needed by LHC

- **using same tools: compare easily results** across experiments
 - not only desirable but **necessary for combinations**

RooStats is built on top of the **RooFit toolkit** :

- **data modelling language** (for PDFs, likelihoods, ...)

RooFit Workspaces

G. Schott
PHYSTAT2011

RooWorkspace class of RooFit: possibility to save it to a ROOT file

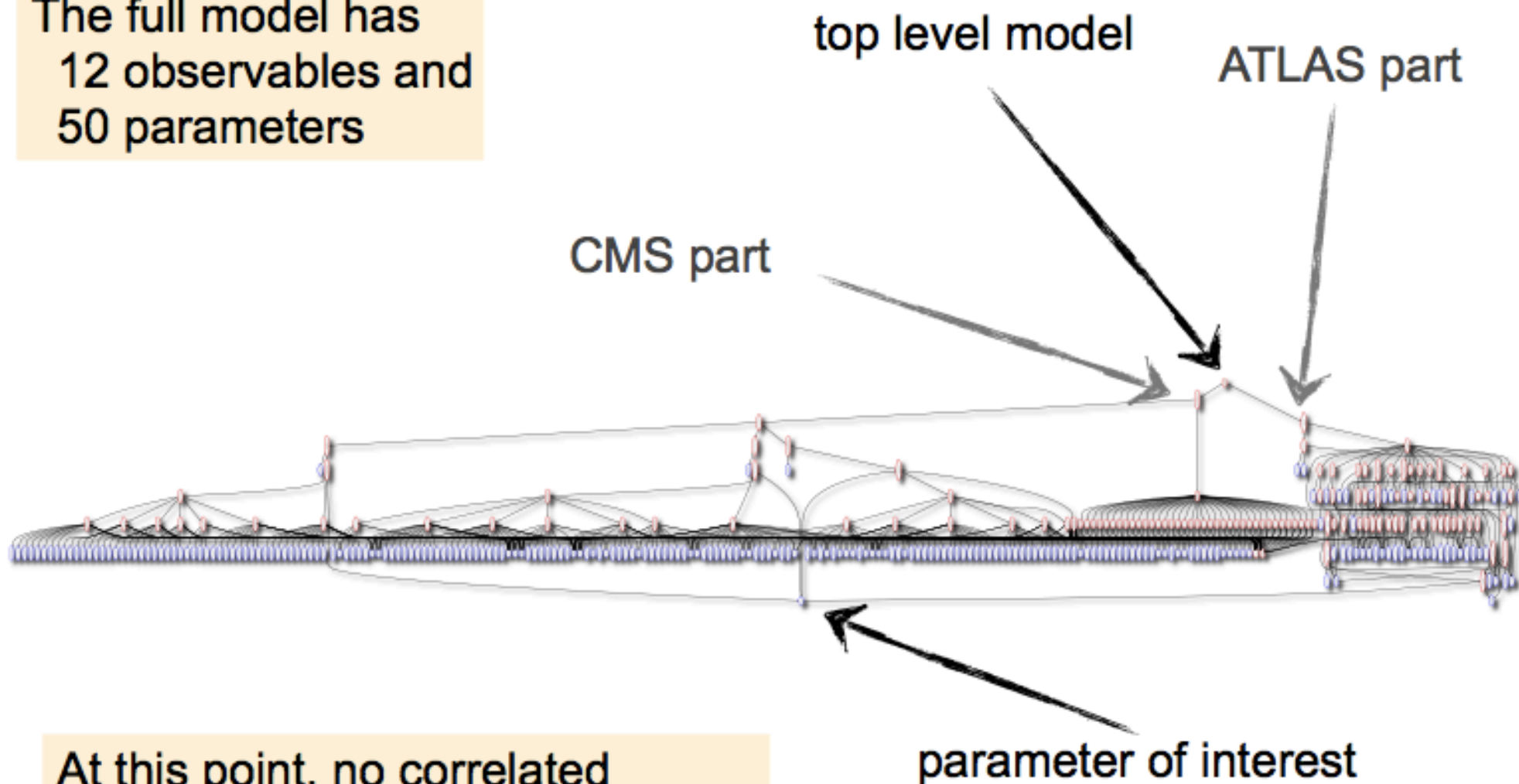
- very good for **electronic publication** of data and likelihood function
- and greatly **help for combination** (that's the format agreed to share between Atlas & CMS)

```
RooWorkspace w("w","joint workspace") ;  
// Import top-level pdfs and all their components, variables  
w.import("channelA.root:w:pdfA",RenameAllVariablesExcept("A","mhiggs"))  
w.import("channelB.root:w:pdfB",RenameVariable("mH","mhiggs")) ;  
w.import("channelC.root:w:pdfC") ;  
// Construct joint pdf  
w.factory("SIMUL::joint(chan[A,B,C],A-pdfA,B-pdfB,C-pdfC)") ;
```

Able to construct full likelihood for combination of channels (or experiments).

Combined ATLAS/CMS Higgs search

The full model has
12 observables and
50 parameters



At this point, no correlated
systematics across experiments

$$\mu = \frac{\sigma BR}{\sigma_{SM} BR_{SM}}$$

Summary (1)

Asymptotic distributions of profile LR applied to an LHC search.

Wilks: $f(q_\mu|\mu)$ for p -value of μ .

Wald approximation for $f(q_\mu|\mu')$.

“Asimov” data set used to estimate median q_μ for sensitivity.

Gives σ of distribution of estimator of μ .

Asymptotic formulae especially useful for estimating sensitivity in high-dimensional parameter space.

Can always check with MC for very low data samples and/or when precision crucial.

Summary (2)

Progress on related issues for LHC discovery:

Look elsewhere effect (Gross and Vitells)

CLs problem → Power Constrained Limits (ongoing)

New software for combinations (and other things): RooStats

Needed:

More work on how to parametrize models so as to include a level of flexibility commensurate with the real systematic uncertainty, together with ideas on how to constrain this flexibility experimentally (control measurements).

Extra slides

Profile likelihood ratio for unified interval

We can also use directly

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

as a test statistic for a hypothesized μ .

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

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

Distribution of t_μ

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

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

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

The p -value for an observed value of t_μ is

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

and the corresponding significance is

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

Confidence intervals by inverting a test

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

Specify values of the data that are ‘disfavoured’ by θ (critical region) such that $P(\text{data in critical region}) \leq \gamma$ for a prespecified γ , e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ .

Now **invert** the test to define a **confidence interval** as:

set of θ values that would **not** be rejected in a test of size γ (confidence level is $1 - \gamma$).

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

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

Relation between confidence interval and p -value

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

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

The confidence interval at $CL = 1 - \gamma$ consists of those values of θ that are not rejected.

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

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

Higgs search with profile likelihood

Combination of Higgs boson search channels (ATLAS)

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

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

$$H \rightarrow \gamma\gamma$$

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

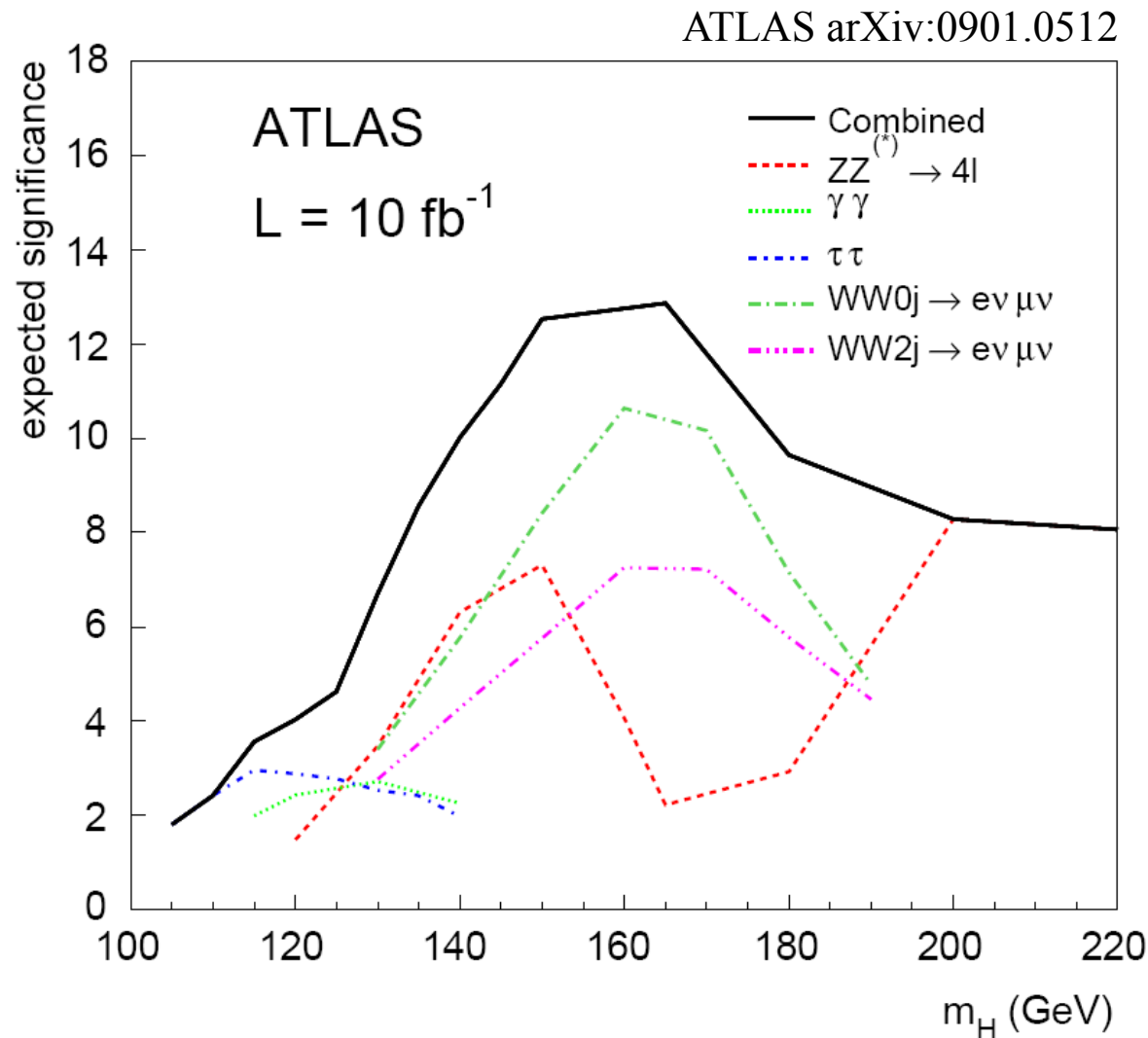
$$H \rightarrow ZZ^{(*)} \rightarrow 4l \quad (l = e, \mu)$$

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

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

Combined median significance



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

An example: ATLAS Higgs search

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

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

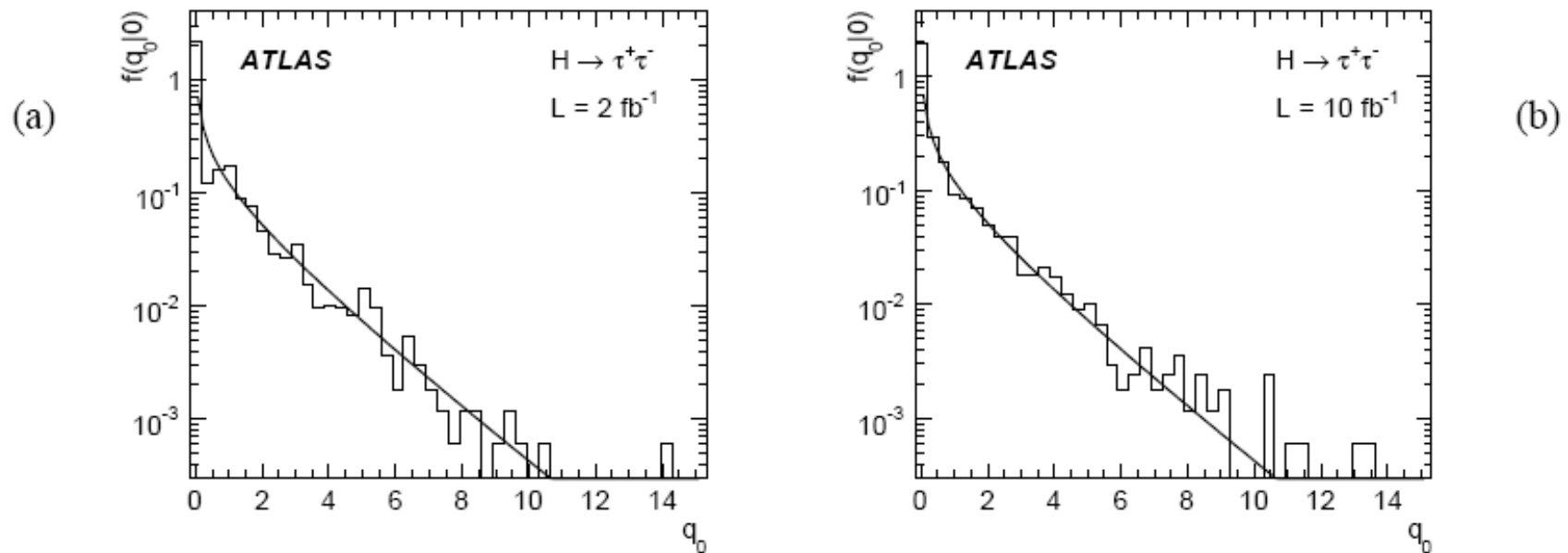
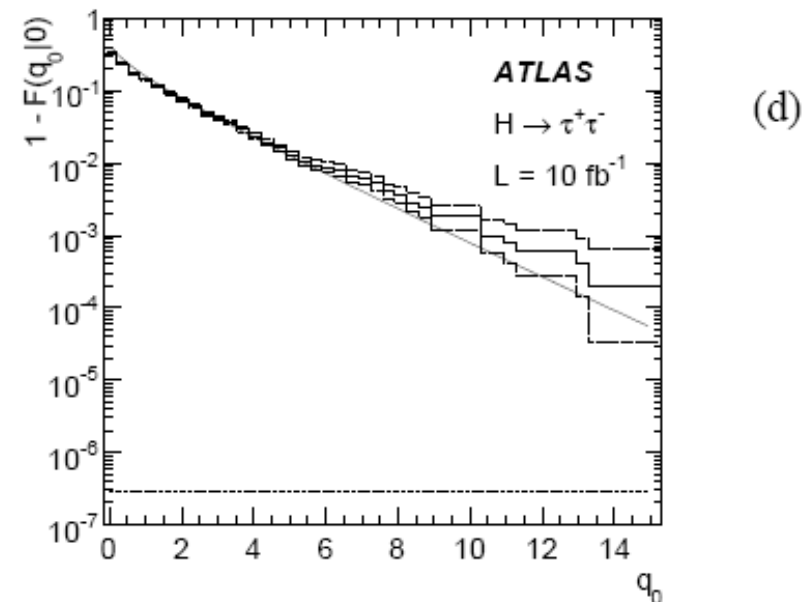
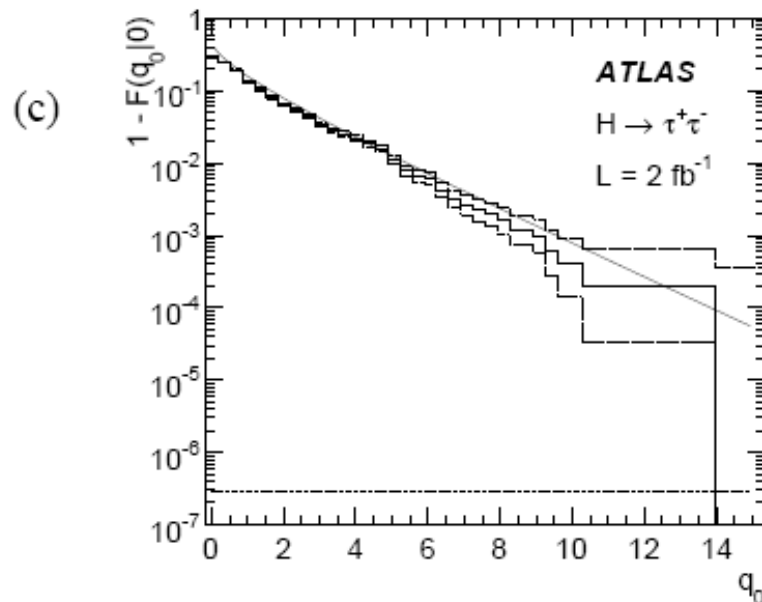


Figure 12: The distribution of the test statistic q_0 for $H \rightarrow \tau^+\tau^-$ under the null background-only hypothesis, for $m_H = 130 \text{ GeV}$ with an integrated luminosity of 2 (a) and 10 (b) fb^{-1} . A $\frac{1}{2}\chi_1^2$ distribution is superimposed. Figures (c) and (d) show $1 -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

Cumulative distributions of q_0

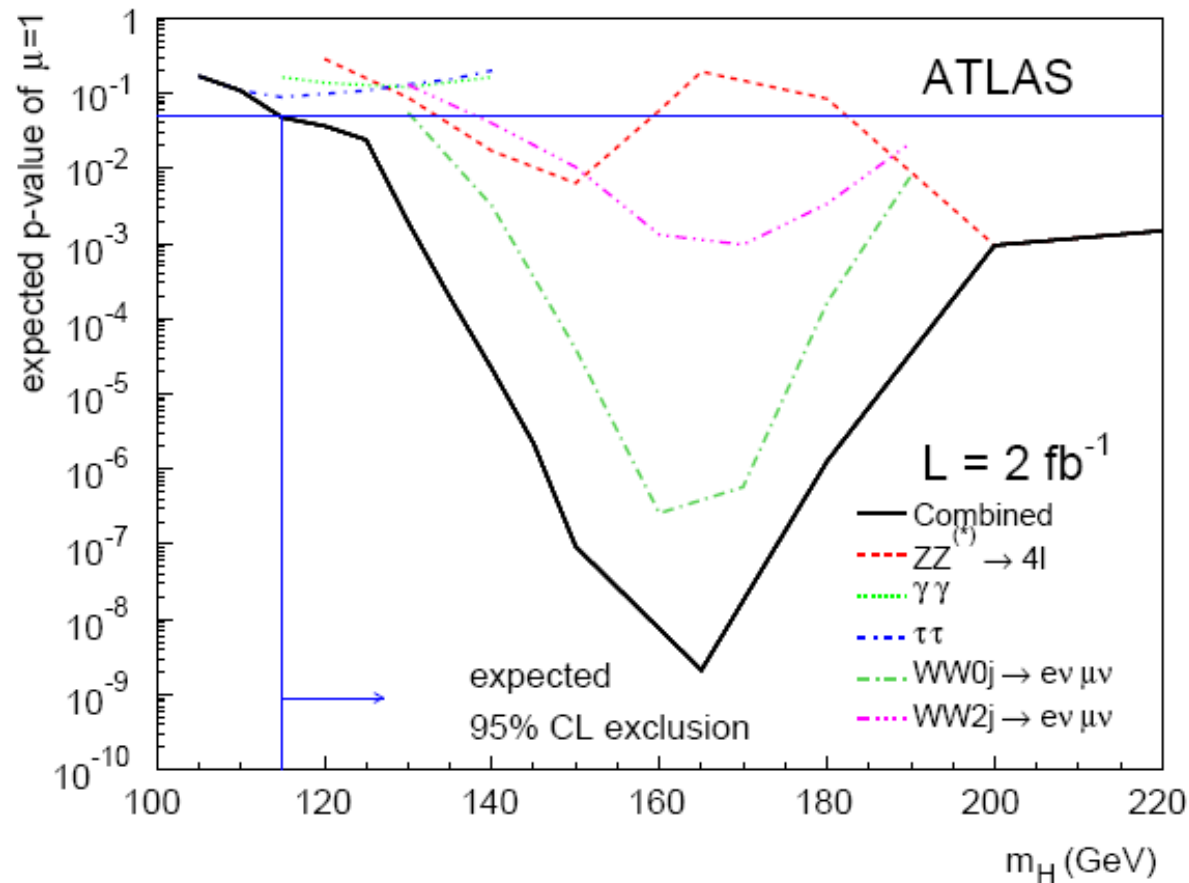
To validate to 5σ level, need distribution out to $q_0 = 25$,
i.e., around 10^8 simulated experiments.

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



Example: exclusion sensitivity

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



Dealing with systematics

S. Caron, G. Cowan, S. Horner, J. Sundermann, E. Gross, 2009 JINST 4 P10009

Suppose one needs to know the shape of a distribution.
Initial model (e.g. MC) is available, but known to be imperfect.

Q: How can one incorporate the systematic error arising from use of the incorrect model?

A: Improve the model.

That is, introduce more adjustable parameters into the model so that for some point in the enlarged parameter space it is very close to the truth.

Then use profile the likelihood with respect to the additional (nuisance) parameters. The correlations with the nuisance parameters will inflate the errors in the parameters of interest.

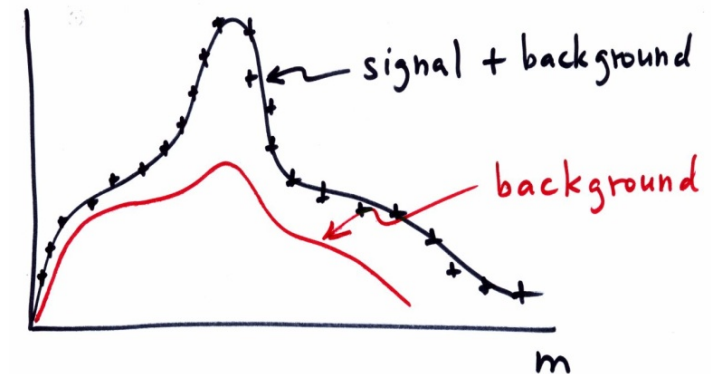
Difficulty is deciding how to introduce the additional parameters.

Example of inserting nuisance parameters

Fit of hadronic mass distribution from a specific τ decay mode.

Important uncertainty in background from non-signal τ modes.

Background rate from other measurements, shape from MC.



Want to include uncertainty in rate, mean, width of background component in a parametric fit of the mass distribution.

Number of events in bin i , $n_i \sim \text{Poisson}(s_i(\boldsymbol{\theta}) + b_i)$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \frac{(s_i(\boldsymbol{\theta}) + b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + b_i)}$$

fit

from MC

Step 1: uncertainty in rate

Scale the predicted background by a factor r : $b_i \rightarrow rb_i$

Uncertainty in r is σ_r

Regard $r_0 = 1$ (“best guess”) as Gaussian (or not, as appropriate) distributed measurement centred about the true value r , which becomes a new “nuisance” parameter in the fit.

New likelihood function is:

$$L(\boldsymbol{\theta}, r) = \prod_{i=1}^N \frac{(s_i(\boldsymbol{\theta}) + rb_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + rb_i)} \frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r-r_0)^2/2\sigma_r^2}$$

For a least-squares fit, equivalent to

$$\chi^2(\boldsymbol{\theta}) \rightarrow \chi^2(\boldsymbol{\theta}) + \frac{(r - r_0)^2}{\sigma_r^2}$$

Dealing with nuisance parameters

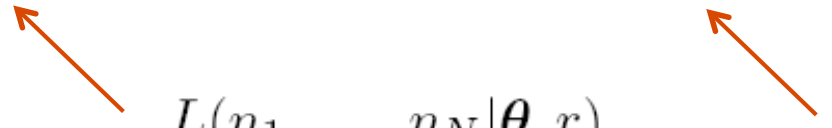
Ways to eliminate the nuisance parameter r from likelihood.

1) Profile likelihood:

$L_p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{r})$, where \hat{r} is value of r that maximizes L for the given $\boldsymbol{\theta}$.

2) Bayesian marginal likelihood:

$$L_m(\boldsymbol{\theta}) = \int \prod_{i=1}^N \frac{(s_i(\boldsymbol{\theta}) + r b_i)^{n_i}}{n_i!} e^{-(s_i(\boldsymbol{\theta}) + r b_i)} \frac{1}{\sqrt{2\pi}\sigma_r} e^{-(r-r_0)^2/2\sigma_r^2} dr$$


 $L(n_1, \dots, n_N | \boldsymbol{\theta}, r)$ $\pi(r)$ (prior)

Profile and marginal likelihoods usually very similar.

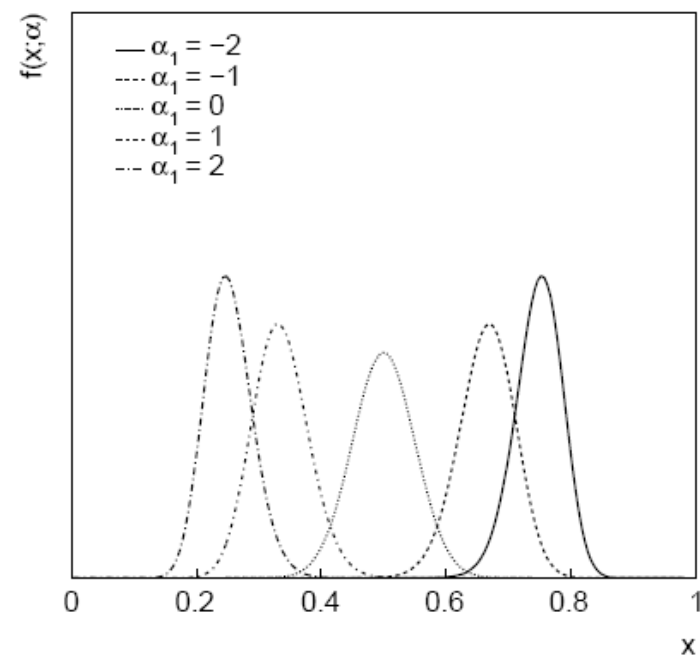
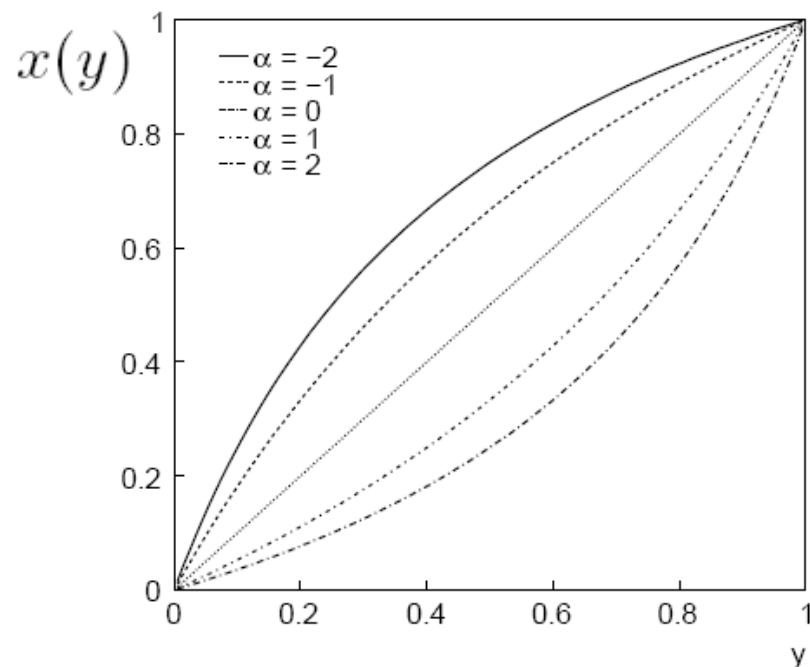
Both are broadened relative to original, reflecting the uncertainty connected with the nuisance parameter.

Step 2: uncertainty in shape

Key is to insert additional nuisance parameters into the model.

E.g. consider a distribution $g(y)$. Let $y \rightarrow x(y)$,

$$x(y) = \begin{cases} \frac{y}{1+\alpha(1-y)} & \alpha \geq 0, \\ \frac{(1-\alpha)y}{1-\alpha y} & \alpha < 0. \end{cases} \quad f(x) = g(y(x)) \left| \frac{dy}{dx} \right|$$

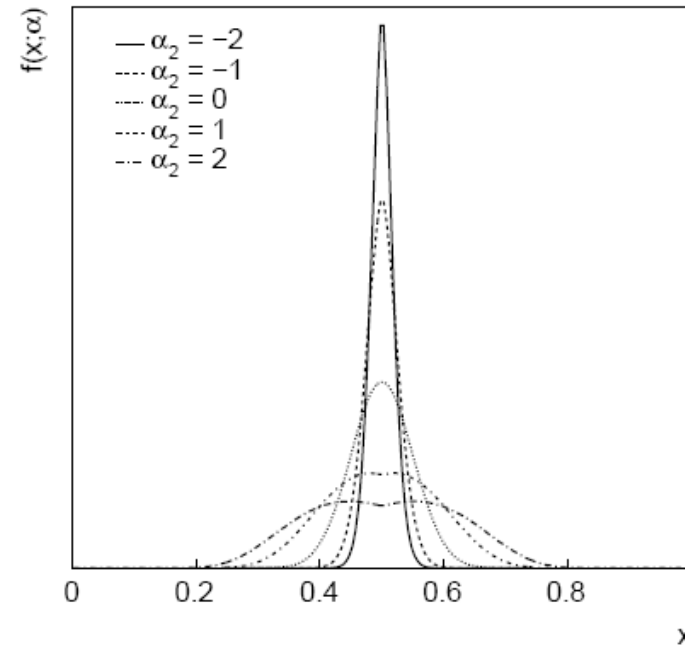
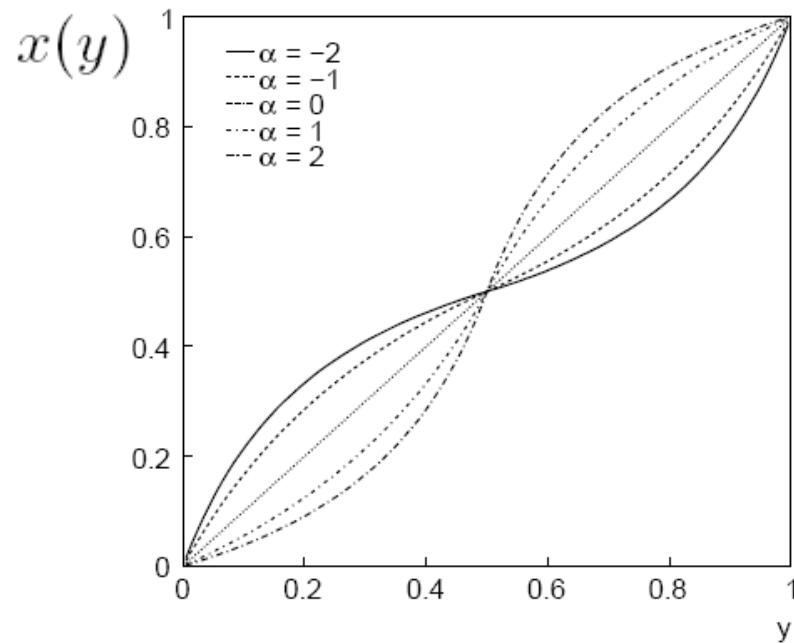


More uncertainty in shape

The transformation can be applied to a spline of original MC histogram (which has shape uncertainty).

Continuous parameter α shifts distribution right/left.

Can play similar game with width (or higher moments), e.g.,



A sample fit (no systematic error)

Consider a Gaussian signal, polynomial background, and also a peaking background whose form is taken from MC:

True mean/width of signal:

$$\mu_s = 0.5, \sigma_s = 0.1$$

True mean/width of background from MC:

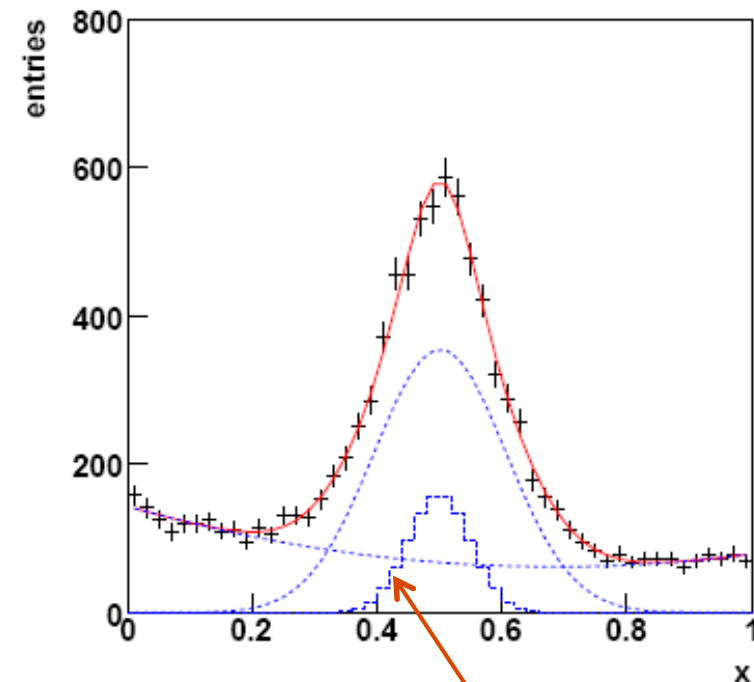
$$\mu_b = 0.5, \sigma_b = 0.05$$

Fit result:

$$\hat{\mu}_s = 0.50025 \pm 0.00232$$

$$\hat{\sigma}_s = 0.10578 \pm 0.00325$$

$$\chi^2 = 30.6 \text{ with } 44 \text{ degrees of freedom}$$



Template
from MC

Sample fit with systematic error

Suppose now the MC template for the peaking background was systematically wrong, having

$$\mu_b = 0.45, \sigma_b = 0.045$$

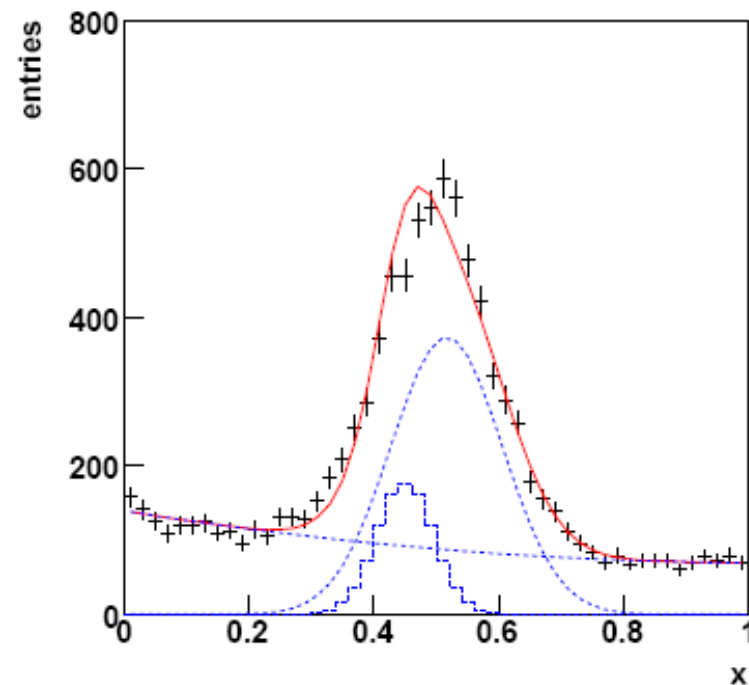
Now fitted values of signal parameters wrong,
poor goodness-of-fit:

$$\hat{\mu}_s = 0.51676 \pm 0.00226$$

$$\hat{\sigma}_s = 0.08933 \pm 0.00308$$

$$\chi^2 = 91.2 \text{ for } 44$$

degrees of freedom



Sample fit with adjustable mean/width


Suppose one regards peak position and width of MC template to have systematic uncertainties:

$$\sigma_{\mu_b} = 0.05 \qquad \sigma_{\sigma_b} = 0.005$$

Incorporate this by regarding the nominal mean/width of the MC template as measurements, so in LS fit add to χ^2 a term:

altered mean
of MC template

original mean
of MC template


$$\left(\frac{\mu_b(\boldsymbol{\alpha}) - \mu_b(0)}{\sigma_{\mu_b}} \right)^2 + \left(\frac{\sigma_b(\boldsymbol{\alpha}) - \sigma_b(0)}{\sigma_{\sigma_b}} \right)^2$$

Sample fit with adjustable mean/width (II)

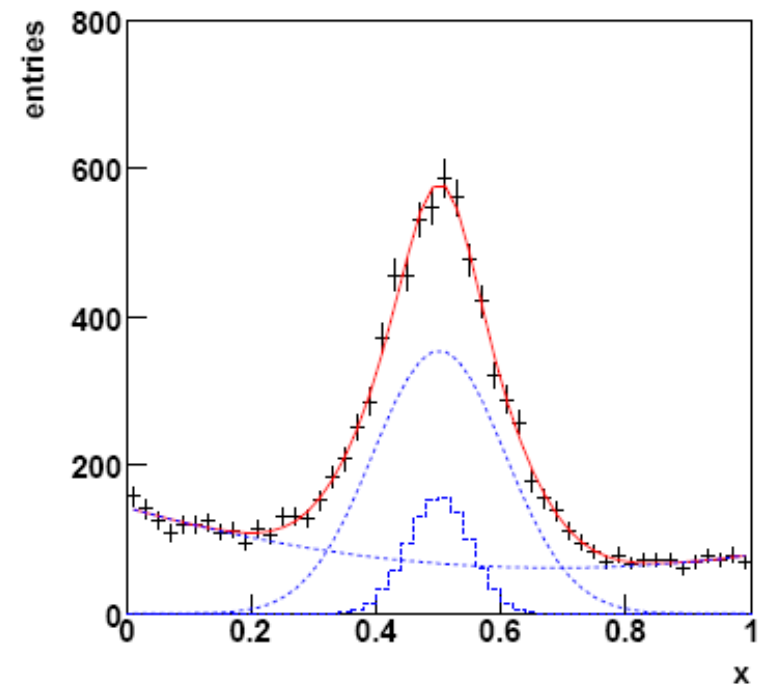
Result of fit is now “good”:

$$\hat{\mu}_s = 0.50014 \pm 0.00290$$

$$\hat{\sigma}_s = 0.10582 \pm 0.00347$$

$$\chi^2 = 32.1 \text{ for } 44$$

degrees of freedom



In principle, continue to add nuisance parameters until data are well described by the model.

Systematic error converted to statistical

One can regard the quadratic difference between the statistical errors with and without the additional nuisance parameters as the contribution from the systematic uncertainty in the MC template:

$$\sigma_{\hat{\mu},\text{sys}} = \sqrt{0.00290^2 - 0.00226^2} = 0.00182$$

$$\sigma_{\hat{\sigma},\text{sys}} = \sqrt{0.00347^2 - 0.00308^2} = 0.00160$$

Formally this part of error has been converted to part of statistical error (because the extended model is \sim correct!).

Systematic error from “shift method”

Note that the systematic error regarded as part of the new statistical error (previous slide) is much smaller than the change one would find by simply “shifting” the templates plus/minus one standard deviation, holding them constant, and redoing the fit. This gives:

$$\Delta\hat{\mu}_{\text{sys}} = |0.50025 - 0.51676| = 0.01651$$

$$\Delta\hat{\sigma}_{\text{sys}} = |0.10578 - 0.08933| = 0.01645$$

This is not necessarily “wrong”, since here we are not improving the model by including new parameters.

But in any case it’s best to improve the model!

Issues with finding an improved model

Sometimes, e.g., if the data set is very large, the total χ^2 can be very high (bad), even though the absolute deviation between model and data may be small.

It may be that including additional parameters "spoils" the parameter of interest and/or leads to an unphysical fit result well before it succeeds in improving the overall goodness-of-fit.

Include new parameters in a clever (physically motivated, local) way, so that it affects only the required regions.

Use Bayesian approach -- assign priors to the new nuisance parameters that constrain them from moving too far (or use equivalent frequentist penalty terms in likelihood).

Unfortunately these solutions may not be practical and one may be forced to use ad hoc recipes (last resort).