

# Statistical Methods for Particle Physics

## Lecture 3: further topics

<https://indico.ihep.ac.cn/event/9142/>



Weihai High-Energy  
Physics School (WHEPS)  
21-28 August 2019



Glen Cowan  
Physics Department  
Royal Holloway, University of London  
[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)  
[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

## Lecture 1: Introduction and review of fundamentals

Review of probability

Parameter estimation, maximum likelihood

Statistical tests for discovery and limits

## Lecture 2: Multivariate methods

Neyman-Pearson lemma

Fisher discriminant, neural networks

Boosted decision trees

## → Lecture 3: Further topics

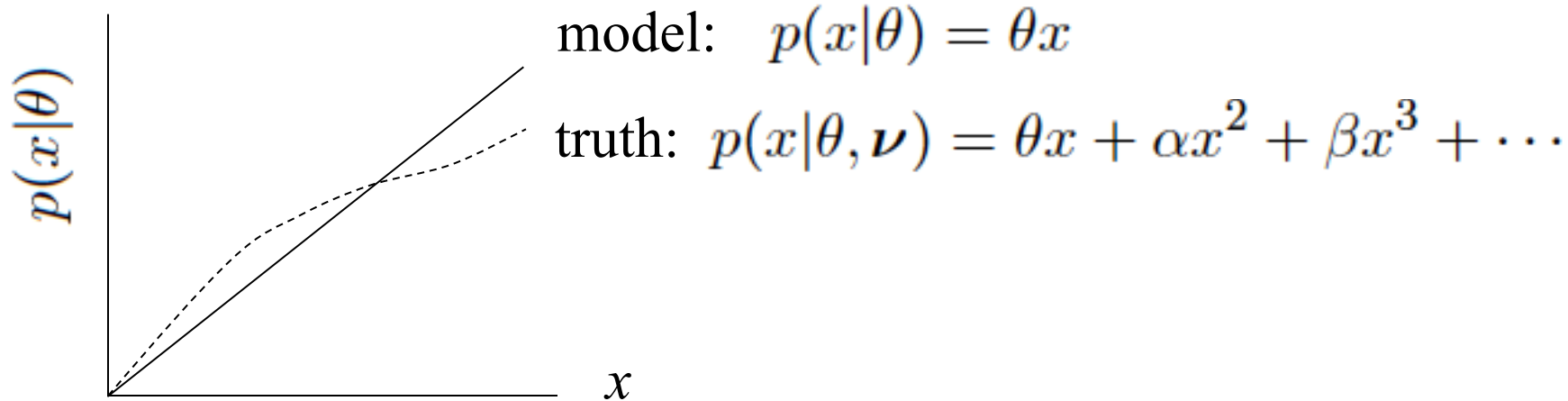
Nuisance parameters (Bayesian and frequentist)

Experimental sensitivity

Revisiting limits

# Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$p(x|\theta) \rightarrow p(x|\theta, \nu)$$

Nuisance parameter  $\leftrightarrow$  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).

# Example: fitting a straight line

Data:  $(x_i, y_i, \sigma_i)$ ,  $i = 1, \dots, n$ .

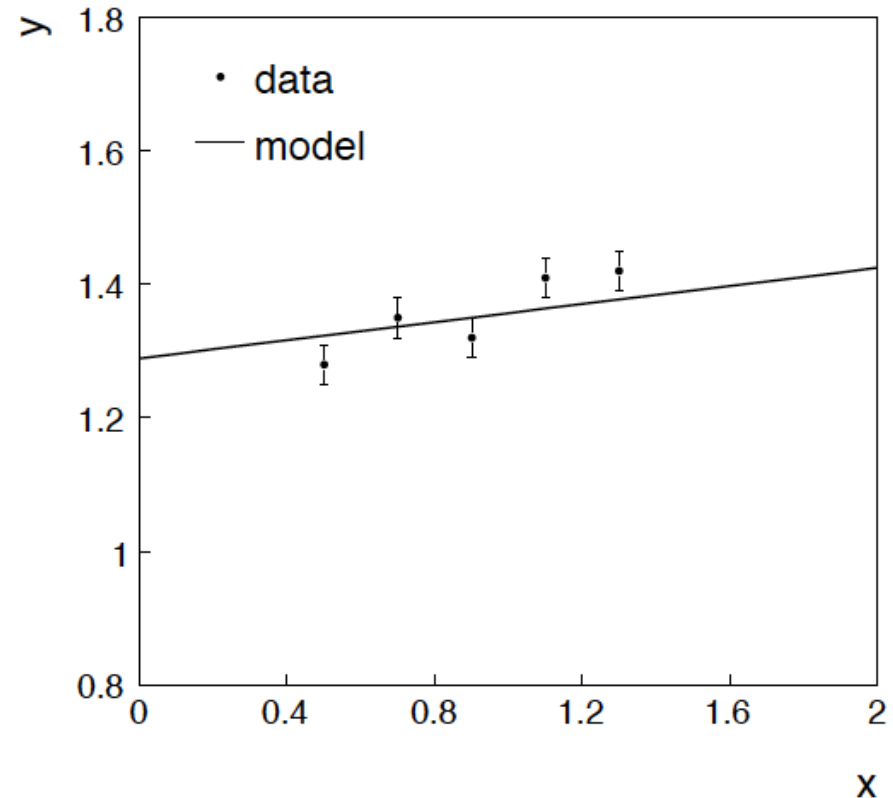
Model:  $y_i$  independent and all follow  $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,$$

assume  $x_i$  and  $\sigma_i$  known.

Goal: estimate  $\theta_0$

Here suppose we don't care about  $\theta_1$  (example of a “nuisance parameter”)



# Maximum likelihood fit with Gaussian data

In this example, the  $y_i$  are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

# $\theta_1$ known a priori

$$L(\theta_0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[ -\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right].$$

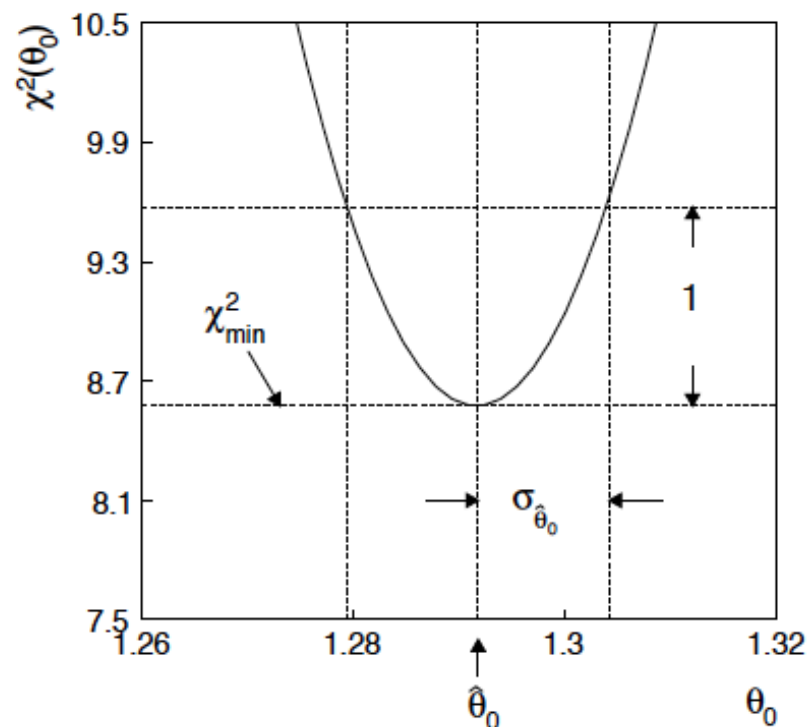
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

For Gaussian  $y_i$ , ML same as LS

Minimize  $\chi^2 \rightarrow$  estimator  $\hat{\theta}_0$ .

Come up one unit from  $\chi_{\min}^2$

to find  $\sigma_{\hat{\theta}_0}$ .



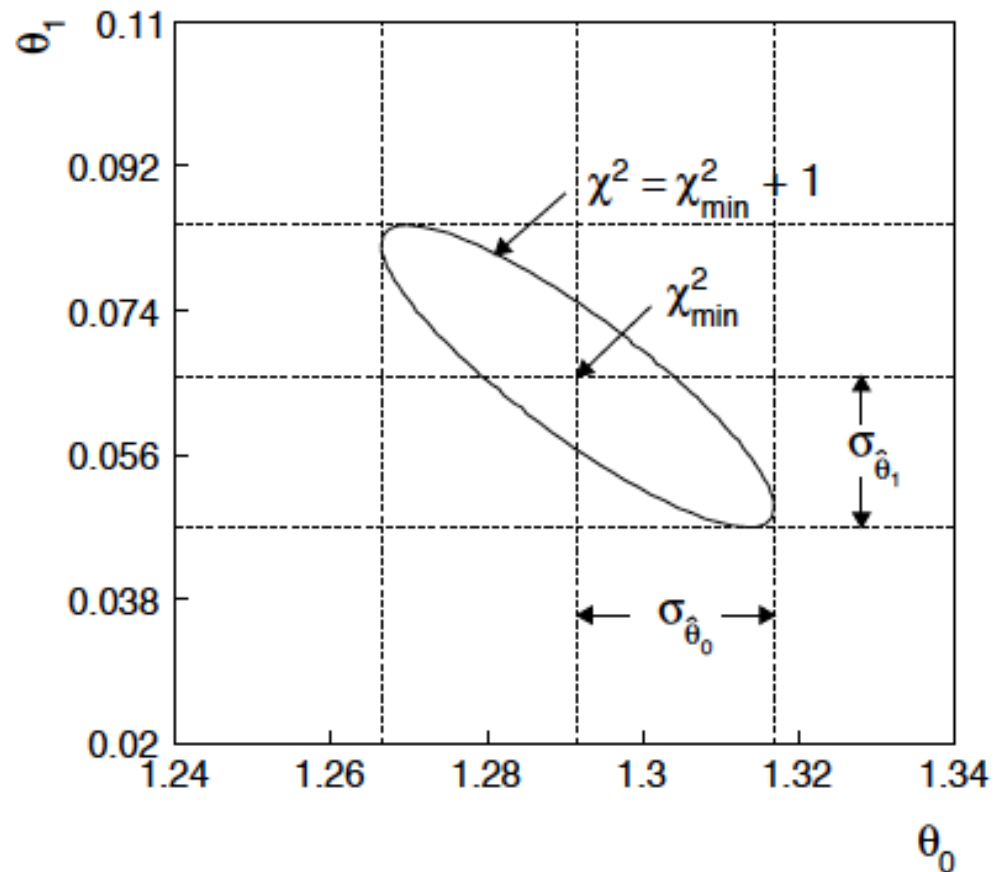
# ML (or LS) fit of $\theta_0$ and $\theta_1$

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from  
tangent lines to contour

$$\chi^2 = \chi_{\min}^2 + 1 .$$

Correlation between  
 $\hat{\theta}_0, \hat{\theta}_1$  causes errors  
to increase.

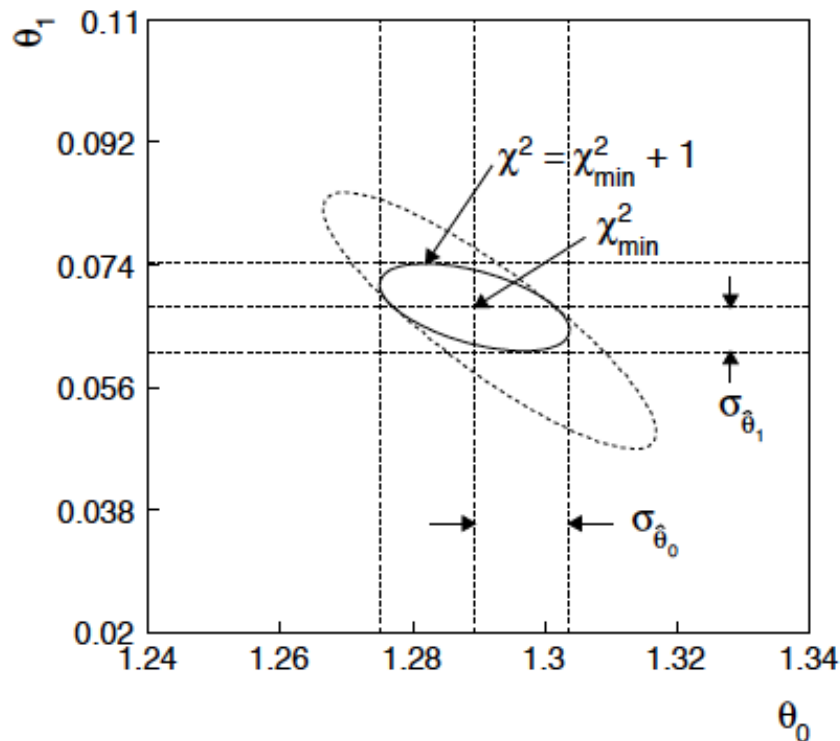


If we have a measurement  $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1 - \theta_1)^2 / 2\sigma_{t_1}^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on  $\theta_1$   
improves accuracy of  $\hat{\theta}_0$ .





# The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value  $\theta$ .

Interpret probability of  $\theta$  as ‘degree of belief’ (subjective).

Need to start with ‘**prior pdf**’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Our experiment has data  $x$ ,  $\rightarrow$  **likelihood function**  $L(x|\theta)$ .

**Bayes’ theorem** tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

**Posterior pdf**  $p(\theta|x)$  contains all our knowledge about  $\theta$ .

# Bayesian method

We need to associate prior probabilities with  $\theta_0$  and  $\theta_1$ , e.g.,

$$\begin{aligned}\pi(\theta_0, \theta_1) &= \pi_0(\theta_0) \pi_1(\theta_1) && \text{'non-informative', in any} \\ \pi_0(\theta_0) &= \text{const.} && \text{case much broader than } L(\theta_0) \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2} && \leftarrow \text{based on previous} \\ &&& \text{measurement}\end{aligned}$$

Putting this into Bayes' theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i; \theta_0, \theta_1))^2 / 2\sigma_i^2} \pi_0 \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2}$$

posterior  $\propto$  likelihood  $\times$  prior

# Bayesian method (continued)

We then integrate (marginalize)  $p(\theta_0, \theta_1 | x)$  to find  $p(\theta_0 | x)$ :

$$p(\theta_0 | x) = \int p(\theta_0, \theta_1 | x) d\theta_1 .$$

In this example we can do the integral (rare). We find

$$p(\theta_0 | x) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2 / 2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 = \text{same as ML estimator}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \text{ (same as before)}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

# Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x) d\theta_1 .$$

often high dimensionality and impossible in closed form,  
also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized  
Bayesian computation.




MCMC (e.g., Metropolis-Hastings algorithm) generates  
**correlated** sequence of random numbers:

cannot use for many applications, e.g., detector MC;  
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional  $\vec{\theta}$  ,  
look, e.g., only at distribution of parameters of interest.

# MCMC basics: Metropolis-Hastings algorithm

Goal: given an  $n$ -dimensional pdf  $p(\vec{\theta})$ ,  
generate a sequence of points  $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

- 1) Start at some point  $\vec{\theta}_0$
- 2) Generate  $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$   Proposal density  $q(\vec{\theta}; \vec{\theta}_0)$   
e.g. Gaussian centred  
about  $\vec{\theta}_0$
- 3) Form Hastings test ratio  $\alpha = \min \left[ 1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$
- 4) Generate  $u \sim \text{Uniform}[0, 1]$
- 5) If  $u \leq \alpha$ ,  $\vec{\theta}_1 = \vec{\theta}$ ,  move to proposed point  
else  $\vec{\theta}_1 = \vec{\theta}_0$   old point repeated
- 6) Iterate

## Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric:  $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

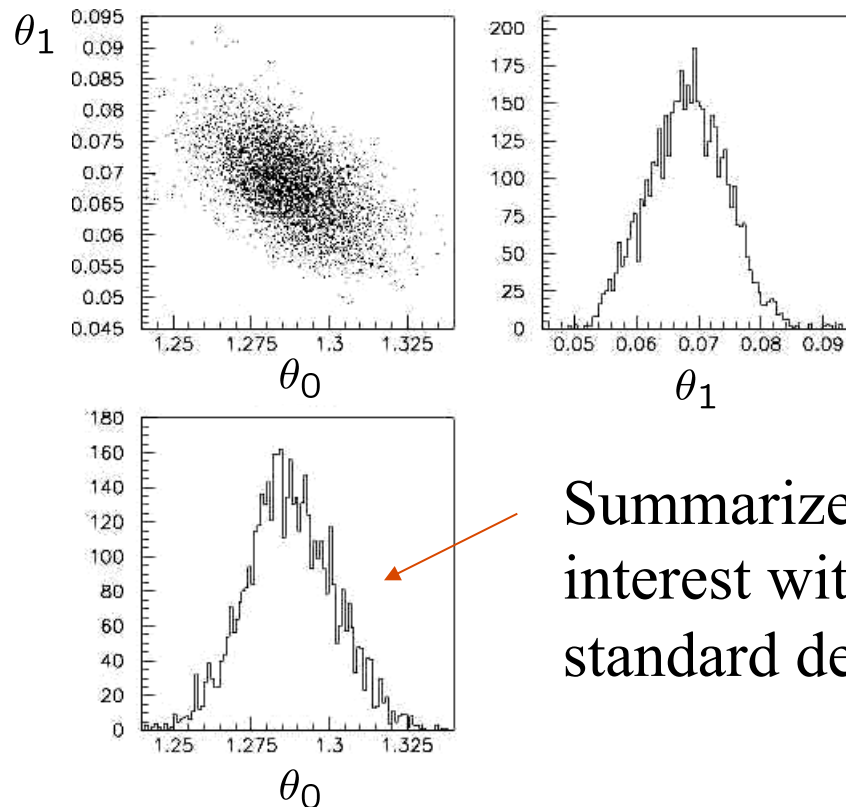
Test ratio is (*Metropolis-Hastings*):  $\alpha = \min \left[ 1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher  $p(\vec{\theta})$ , take it; if not, only take the step with probability  $p(\vec{\theta})/p(\vec{\theta}_0)$ .

If proposed step rejected, hop in place.

# Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

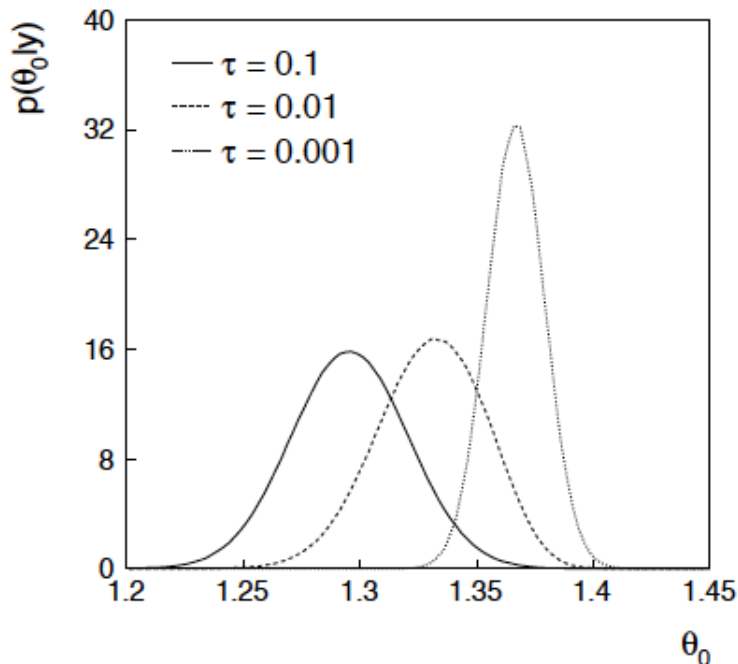
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

# Bayesian method with alternative priors

Suppose we don't have a previous measurement of  $\theta_1$  but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \quad \theta_1 \geq 0, \quad \tau = 0.1 .$$

From this we obtain (numerically) the posterior pdf for  $\theta_0$ :



This summarizes all knowledge about  $\theta_0$ .

Look also at result from variety of priors.



# 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 $b$ , $\sigma_b$ :

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

(b) Profile likelihood ratio test & Asimov:

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

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

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

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

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

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

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

 “profiled” values of  $\boldsymbol{\nu}$

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

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

## 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:

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

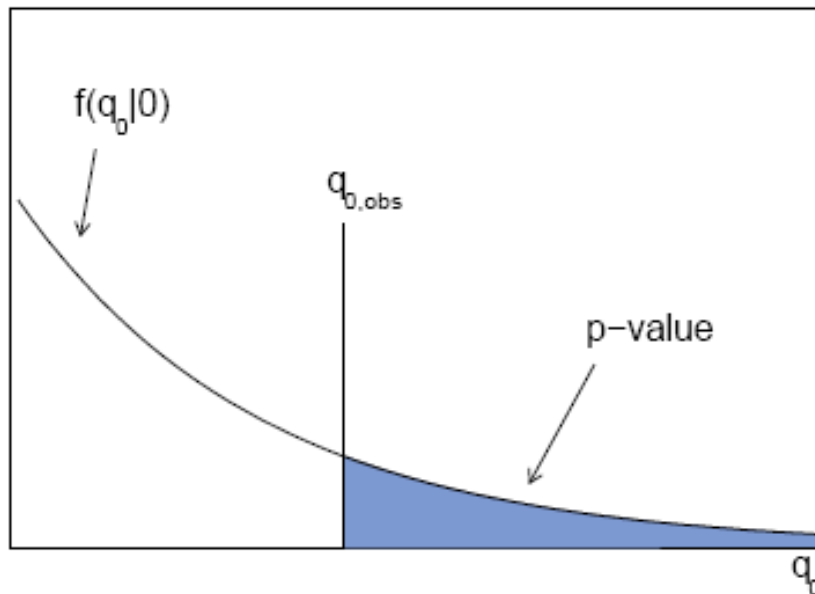
In large sample limit,  $f(q_0|0)$  independent of nuisance parameters;  $f(q_0|\mu')$  depends on nuisance parameters through  $\sigma$ .

## $p$ -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

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

use e.g. asymptotic formula



From  $p$ -value get equivalent significance,

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

## 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

$$F(q_0|0) = \Phi(\sqrt{q_0})$$

The  $p$ -value of the  $\mu = 0$  hypothesis is

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

Therefore the discovery significance  $Z$  is simply

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

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

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

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

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

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



# 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:

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

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

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

# Approximate Poisson significance (continued)

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

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

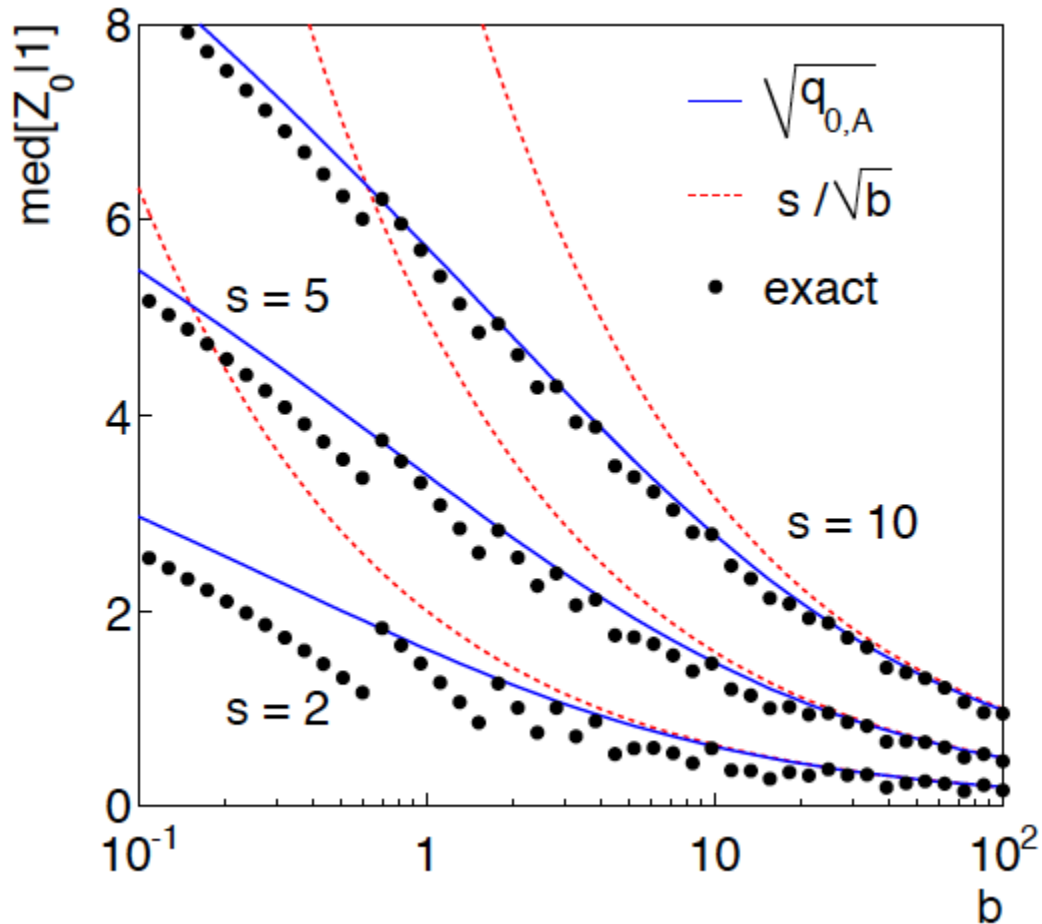
To find  $\text{median}[Z|s]$ , let  $n \rightarrow s + b$  (i.e., the Asimov data set):

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

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

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

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

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

This has been used to optimize some analyses e.g. where  $\sigma_b$  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,  $\tau$  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):

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

# Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

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

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

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

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

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

# 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:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

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

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

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



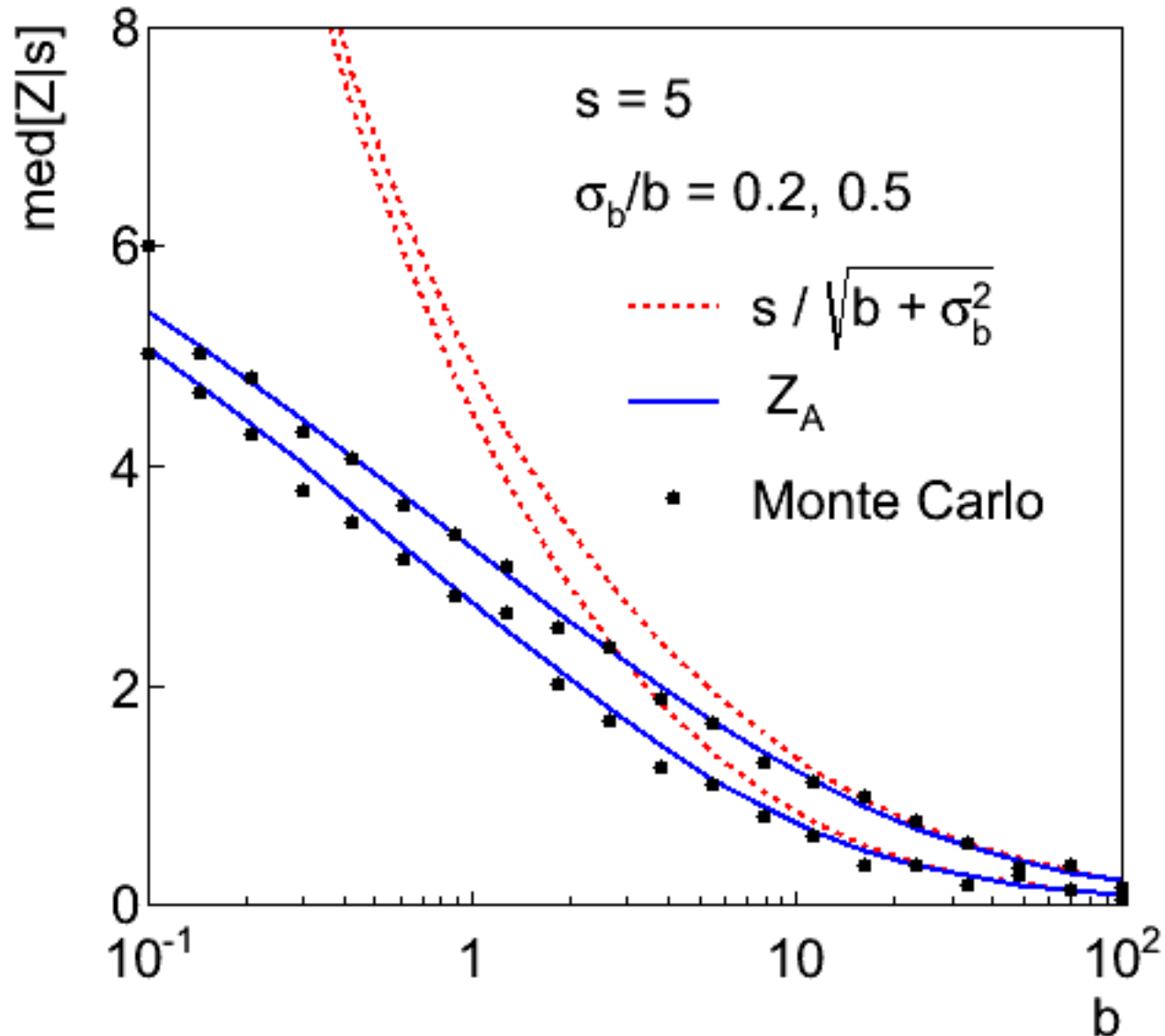
# Limiting cases

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

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}} \left( 1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

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

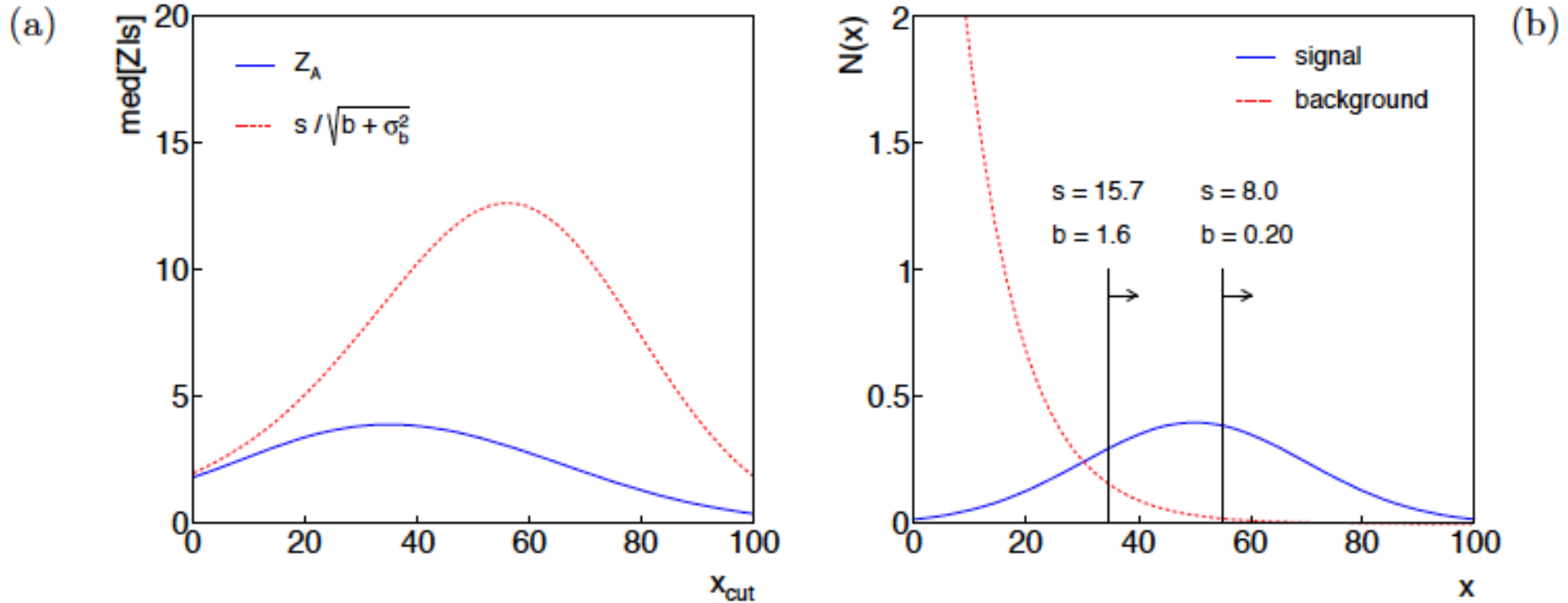


Figure 1: (a) The expected significance as a function of the cut value  $x_{\text{cut}}$ ; (b) the distributions of signal and background with the optimal cut value indicated.

# Summary on discovery sensitivity

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

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

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).

# Return to interval estimation

Suppose a model contains a parameter  $\mu$ ; we want to know which values are consistent with the data and which are disfavoured.

Carry out a test of size  $\alpha$  for all values of  $\mu$ .

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

The probability that the true value of  $\mu$  will be rejected is not greater than  $\alpha$ , so by construction the confidence interval will contain the true value of  $\mu$  with probability  $\geq 1 - \alpha$ .

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

If the test is formulated in terms of a  $p$ -value,  $p_\mu$ , then the confidence interval represents those values of  $\mu$  for which  $p_\mu > \alpha$ .

To find the end points of the interval, set  $p_\mu = \alpha$  and solve for  $\mu$ .

# 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  $\mu$  one can use

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

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized  $\mu$ :

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

Large sample approximation:

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

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

# Monte Carlo test of asymptotic formulae

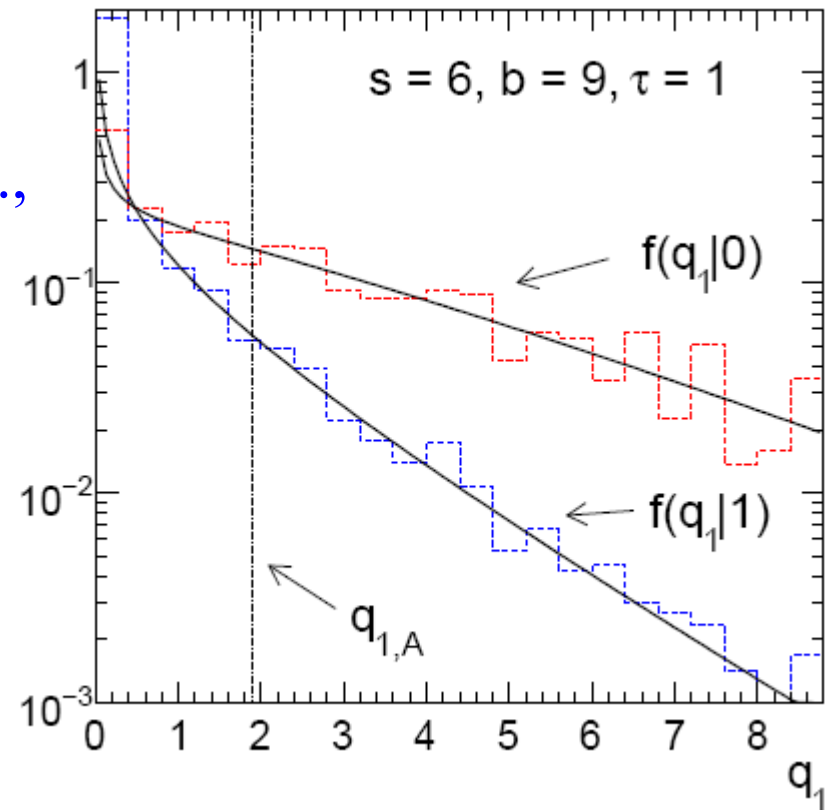
Consider again  $n \sim \text{Poisson}(\mu s + b)$ ,  $m \sim \text{Poisson}(\tau b)$   
 Use  $q_\mu$  to find  $p$ -value of hypothesized  $\mu$  values.

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”.

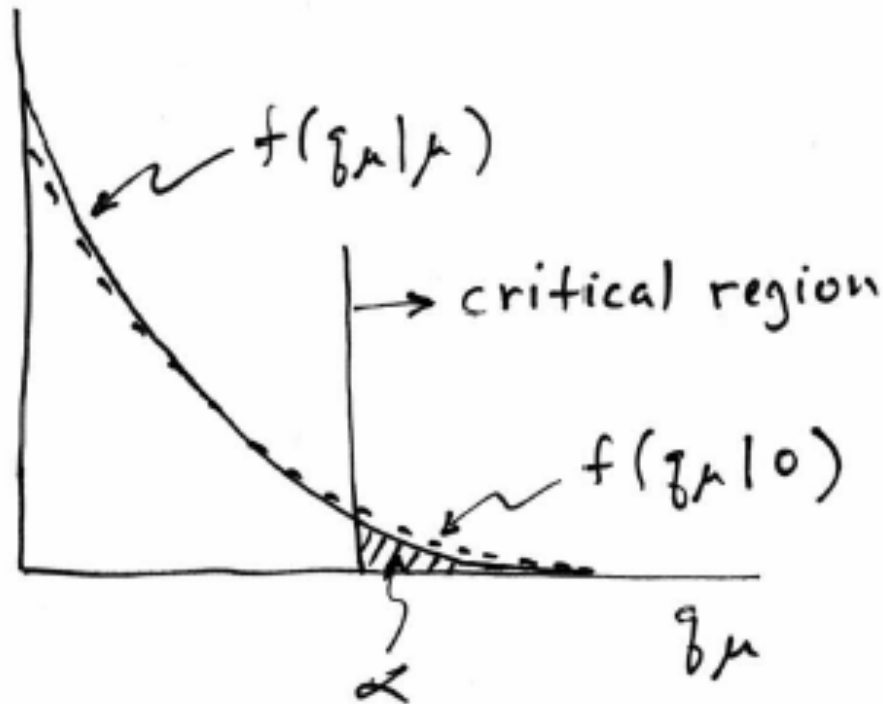
Here asymptotic formulae good  
 for  $s = 6$ ,  $b = 9$ .



## Low sensitivity to $\mu$

It can be that the effect of a given hypothesized  $\mu$  is very small relative to the background-only ( $\mu = 0$ ) prediction.

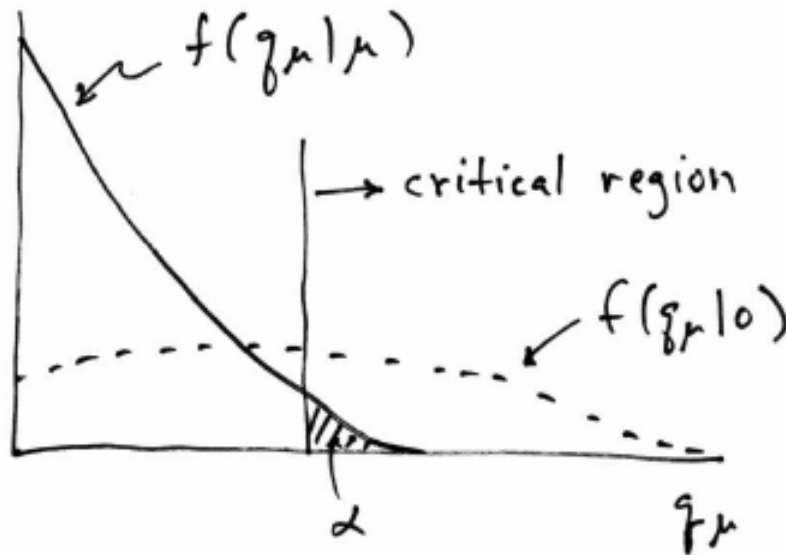
This means that the distributions  $f(q_\mu|\mu)$  and  $f(q_\mu|0)$  will be almost the same:





# Having sufficient sensitivity

In contrast, having sensitivity to  $\mu$  means that the distributions  $f(q_\mu|\mu)$  and  $f(q_\mu|0)$  are more separated:

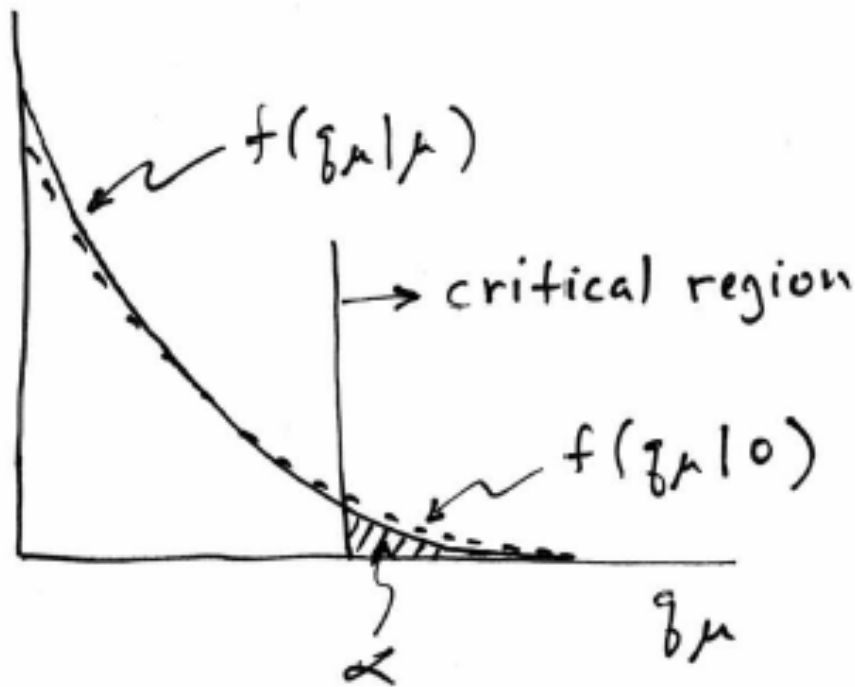


That is, the power (probability to reject  $\mu$  if  $\mu = 0$ ) is substantially higher than  $\alpha$ . Use this power as a measure of the sensitivity.

# Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject  $\mu$  if  $\mu$  is true is  $\alpha$  (e.g., 5%).

And the probability to reject  $\mu$  if  $\mu = 0$  (the power) is only slightly greater than  $\alpha$ .



This means that with probability of around  $\alpha = 5\%$  (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g.,  $m_H = 1000$  TeV).

“Spurious exclusion”

# Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

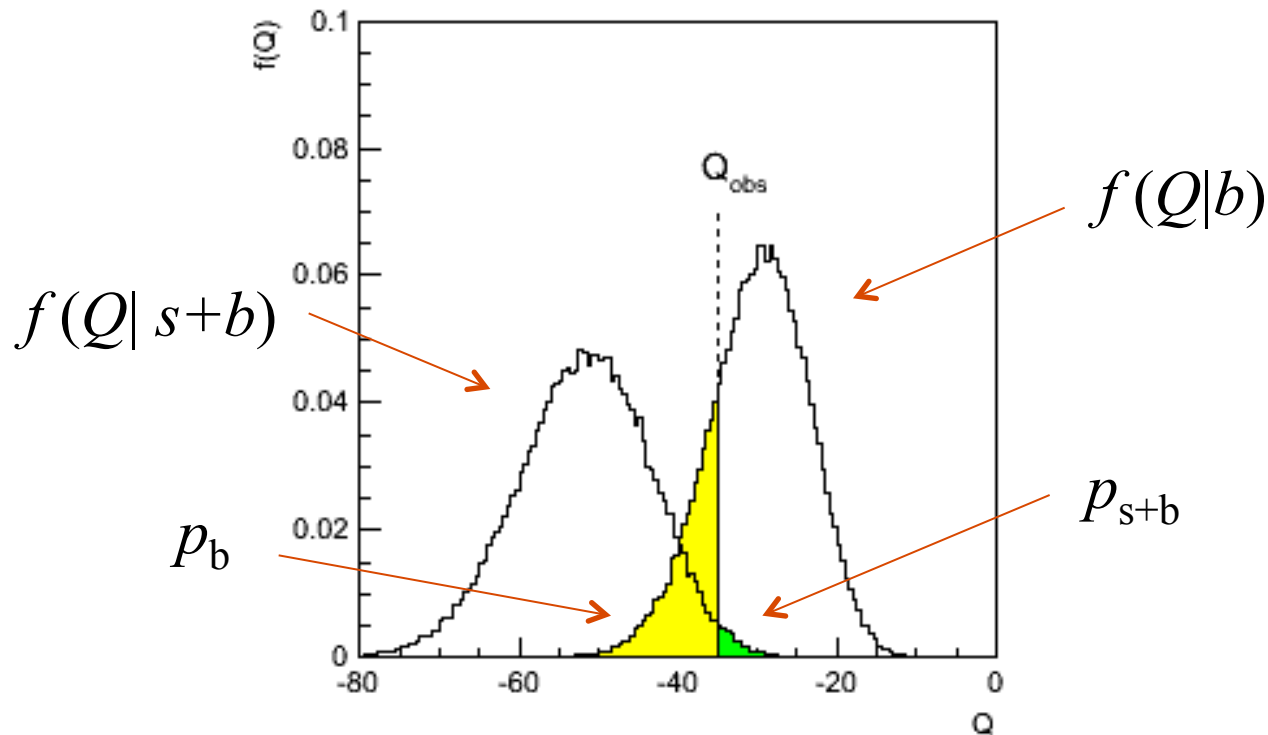
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the “ $CL_s$ ” procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.

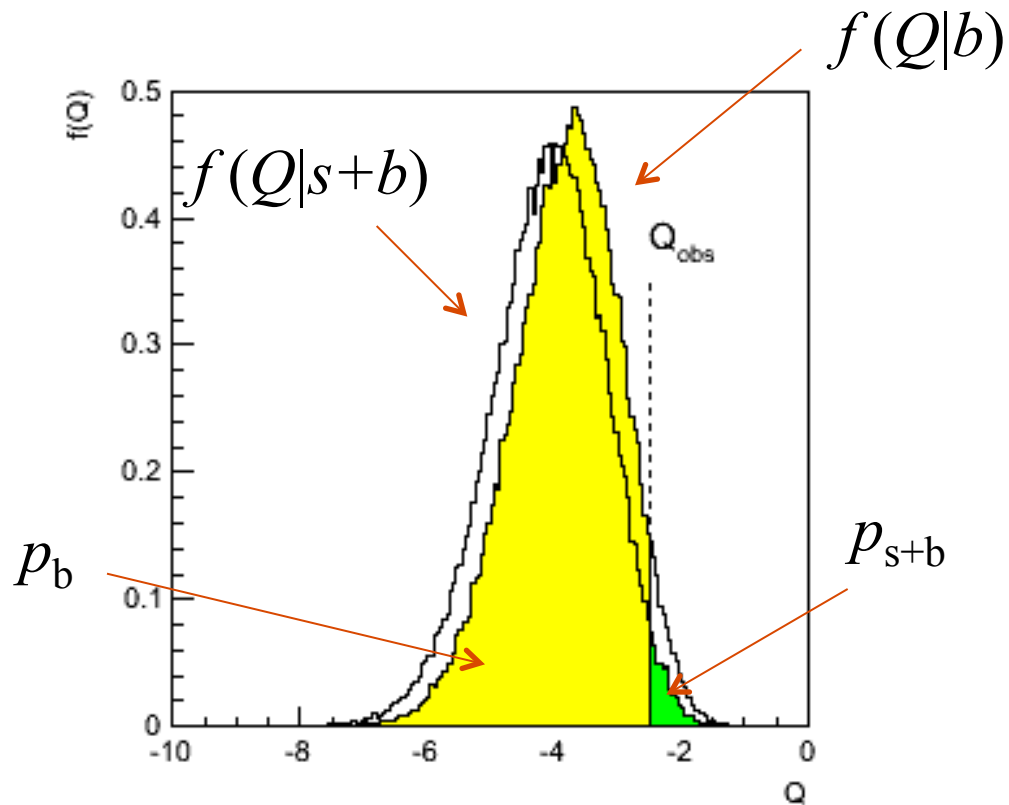
# The $CL_s$ procedure

In the usual formulation of  $CL_s$ , one tests both the  $\mu = 0$  ( $b$ ) and  $\mu > 0$  ( $\mu s + b$ ) hypotheses with the same statistic  $Q = -2 \ln L_{s+b}/L_b$ :



## The $CL_s$ procedure (2)

As before, “low sensitivity” means the distributions of  $Q$  under  $b$  and  $s+b$  are very close:



# The $CL_s$ procedure (3)

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  $CL_b$  ( $\sim$  one minus the  $p$ -value of the  $b$ -only hypothesis), i.e.,

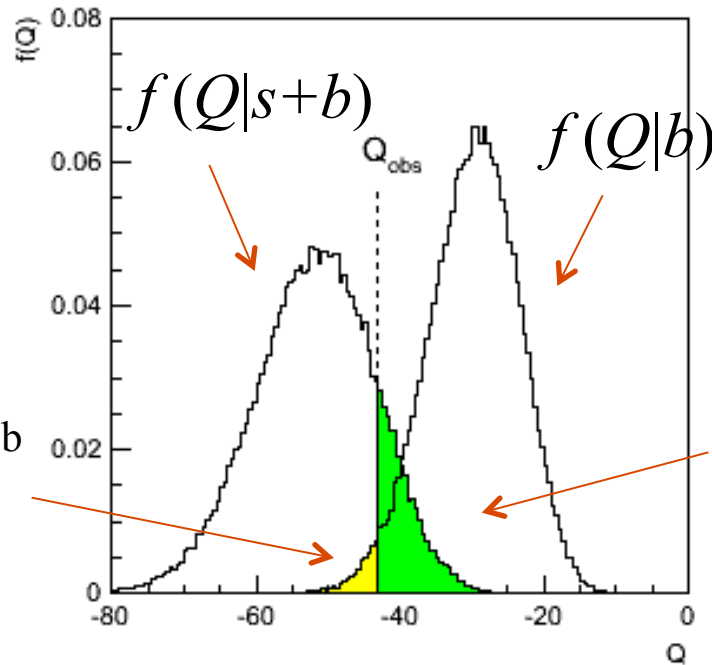
Define:

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

Reject  $s+b$  hypothesis if:

$$CL_s \leq \alpha$$

$$1 - CL_b = p_b$$



$$CL_{s+b} = p_{s+b}$$

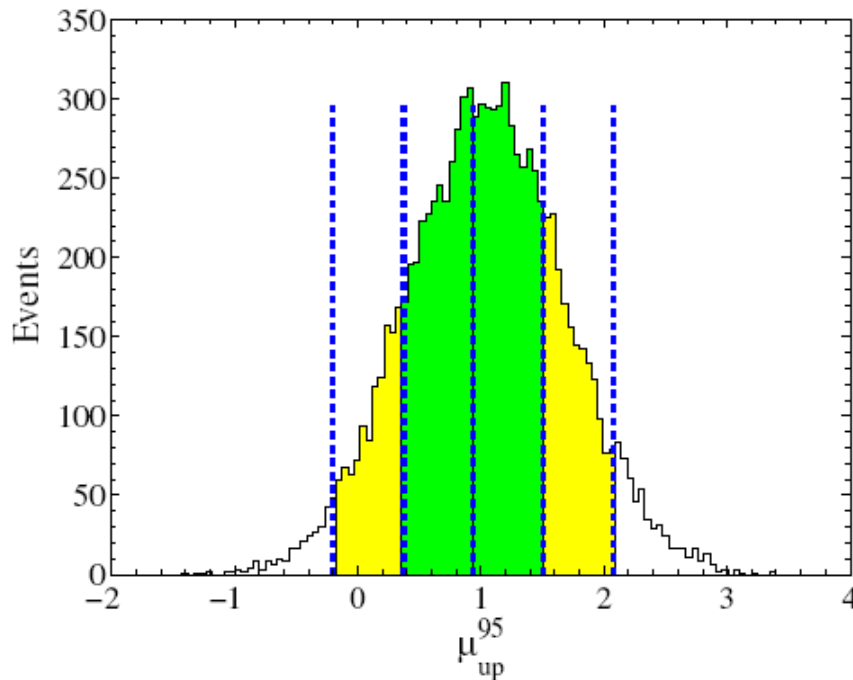
Increases “effective”  $p$ -value when the two distributions become close (prevents exclusion if sensitivity is low).

# Setting upper limits on $\mu = \sigma/\sigma_{\text{SM}}$

Carry out the CLs procedure for the parameter  $\mu = \sigma/\sigma_{\text{SM}}$ , resulting in an upper limit  $\mu_{\text{up}}$ .

In, e.g., a Higgs search, this is done for each value of  $m_{\text{H}}$ .

At a given value of  $m_{\text{H}}$ , we have an observed value of  $\mu_{\text{up}}$ , and we can also find the distribution  $f(\mu_{\text{up}}|0)$ :



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

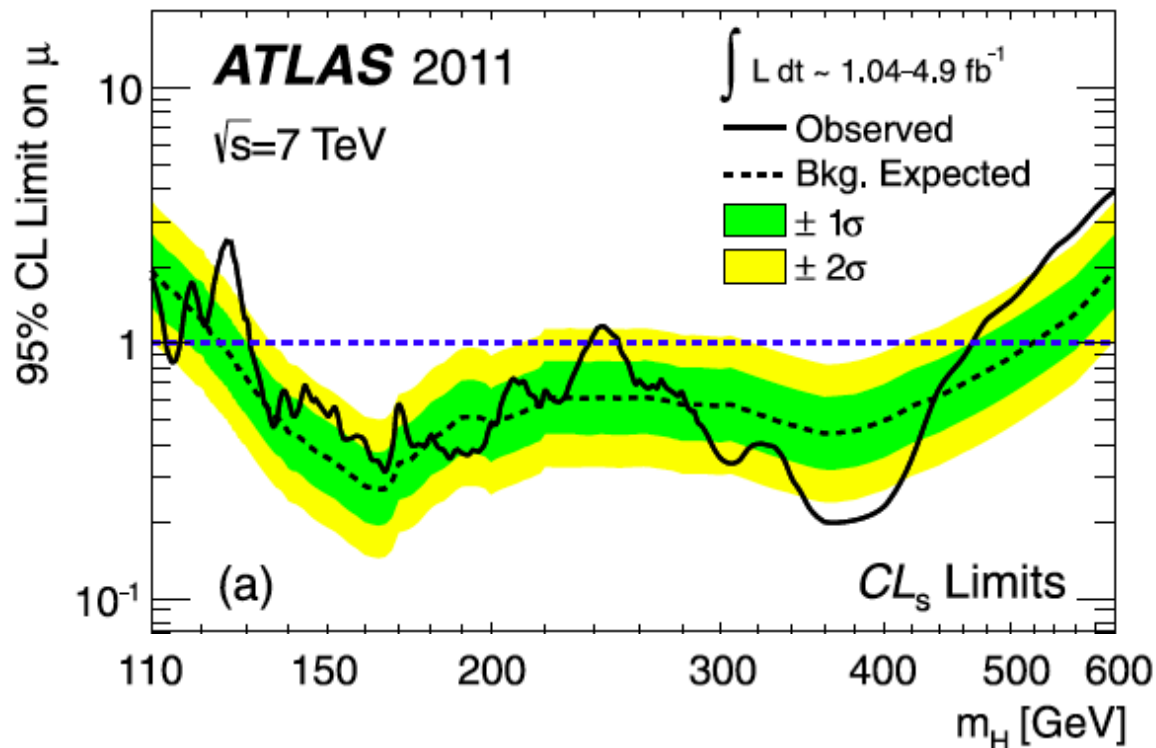
Vertical lines from asymptotic formulae.

# How to read the green and yellow limit plots

For every value of  $m_H$ , find the CLs upper limit on  $\mu$ .

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

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



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



# The Bayesian approach to limits

In Bayesian statistics need to start with ‘prior pdf’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Bayes’ theorem tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf  $p(\theta|x)$  to give interval with any desired probability content.

For e.g.  $n \sim \text{Poisson}(s+b)$ , 95% CL upper limit on  $s$  from

$$0.95 = \int_{-\infty}^{s_{\text{sup}}} p(s|n) ds$$

# Bayesian prior for Poisson parameter

Include knowledge that  $s \geq 0$  by setting prior  $\pi(s) = 0$  for  $s < 0$ .

Could try to reflect ‘prior ignorance’ with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as  $L(s)$  dies off for large  $s$ .

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true  $s$ ).

## Bayesian interval with flat prior for $s$

Solve to find limit  $s_{\text{up}}$ :

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1} [p, 2(n+1)] - b$$

where

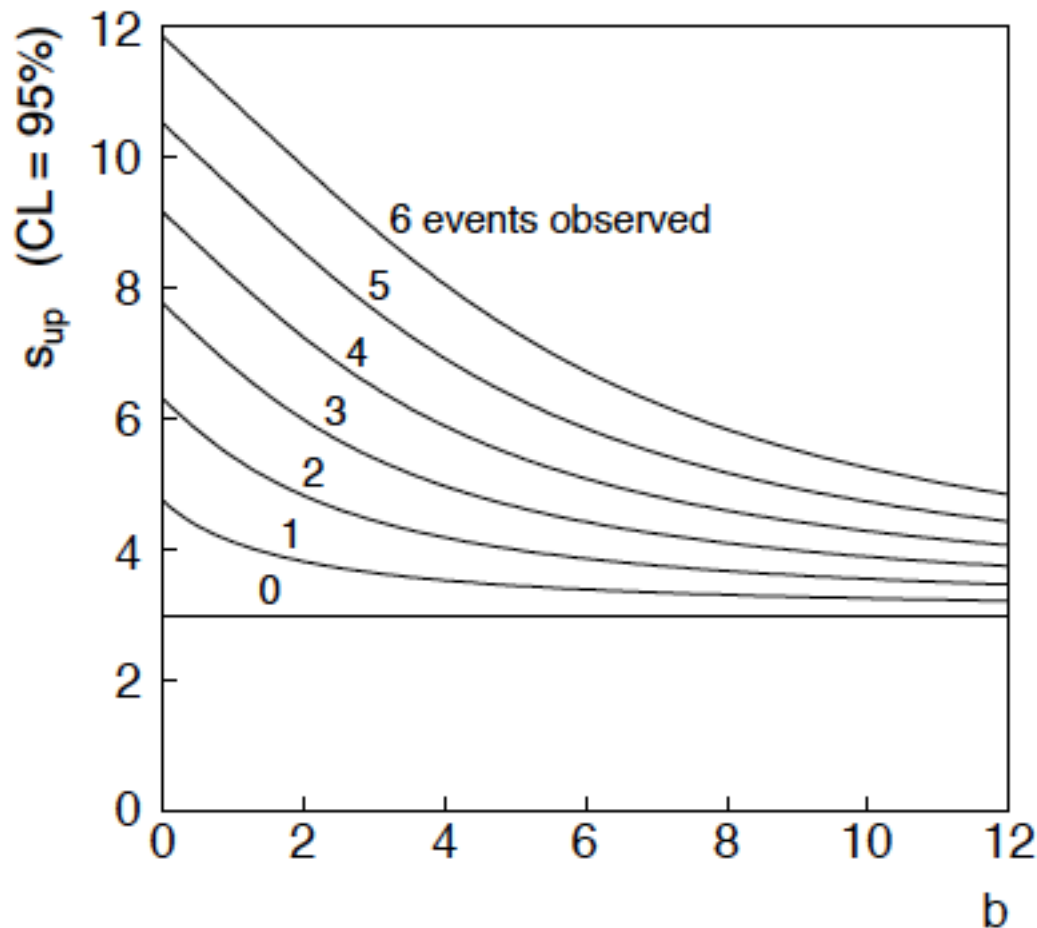
$$p = 1 - \alpha \left( 1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

For special case  $b = 0$ , Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

# Bayesian interval with flat prior for $s$

For  $b > 0$  Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on  $b$  if  $n = 0$ .



# Finally

Three lectures only enough for a brief introduction to:

Statistical tests for discovery and limits

Multivariate methods

Bayesian parameter estimation, MCMC

Experimental sensitivity

No time for many important topics

Properties of estimators (bias, variance)

Bayesian approach to discovery (Bayes factors)

The look-elsewhere effect, etc., etc.

Final thought: once the basic formalism is understood, most of the work focuses on writing down the likelihood, e.g.,  $P(\mathbf{x}|\theta)$ , and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).

# Extra slides

## Choice of test for limits (2)

In some cases  $\mu = 0$  is no longer a relevant alternative and we want to try to exclude  $\mu$  on the grounds that some other measure of incompatibility between it and the data exceeds some threshold.

If the measure of incompatibility is taken to be the likelihood ratio with respect to a two-sided alternative, then the critical region can contain both high and low data values.

→ unified intervals, G. Feldman, R. Cousins,  
Phys. Rev. D 57, 3873–3889 (1998)

The Big Debate is whether to use one-sided or unified intervals in cases where small (or zero) values of the parameter are relevant alternatives. Professional statisticians have voiced support on both sides of the debate.

# Unified (Feldman-Cousins) intervals

We can use directly

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

as a test statistic for a hypothesized  $\mu$ .

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

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

G. Feldman and R.D. Cousins, Phys. Rev. D 57 (1998) 3873.

Lower edge of interval can be at  $\mu = 0$ , depending on data.



## Distribution of $t_\mu$

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  $t_\mu$  is

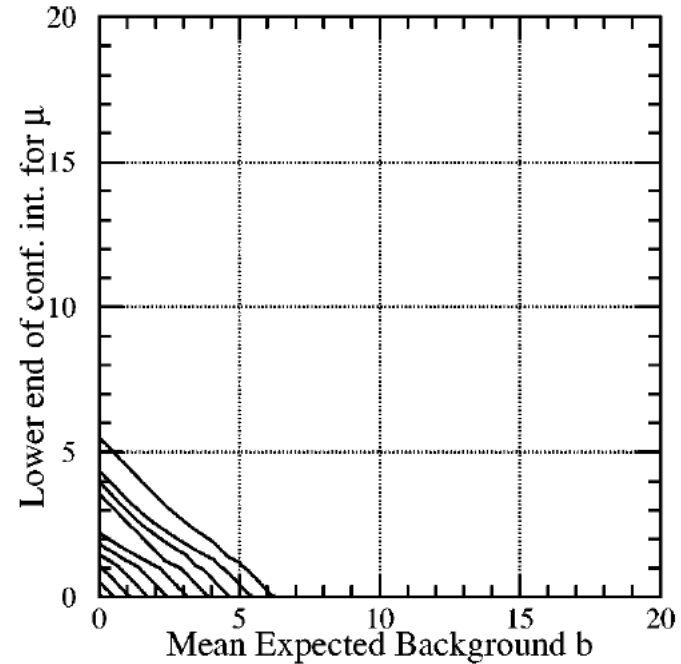
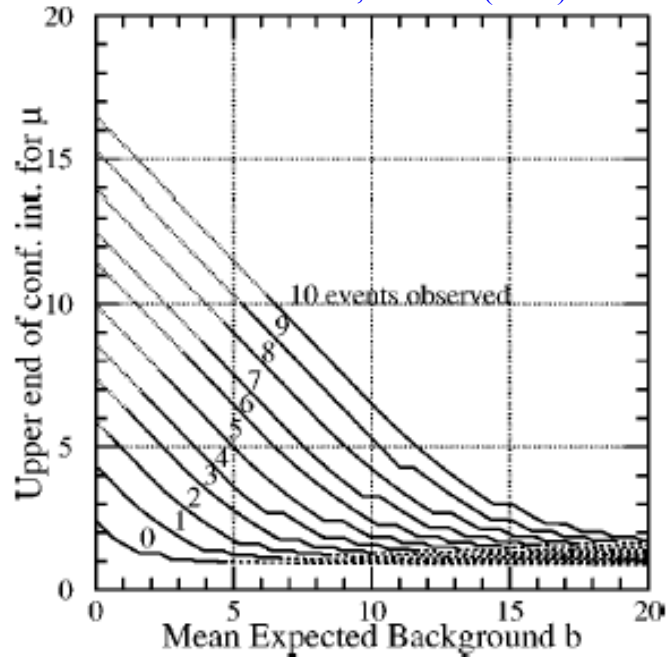
$$p_\mu = 1 - F(t_\mu|\mu) = 2(1 - \Phi(\sqrt{t_\mu}))$$

and the corresponding significance is

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

# Upper/lower edges of F-C interval for $\mu$ versus $b$ for $n \sim \text{Poisson}(\mu+b)$

Feldman & Cousins, PRD 57 (1998) 3873



Lower edge may be at zero, depending on data.

For  $n = 0$ , upper edge has (weak) dependence on  $b$ .

# Feldman-Cousins discussion

The initial motivation for Feldman-Cousins (unified) confidence intervals was to eliminate null intervals.

The F-C limits are based on a likelihood ratio for a test of  $\mu$  with respect to the alternative consisting of all other allowed values of  $\mu$  (not just, say, lower values).

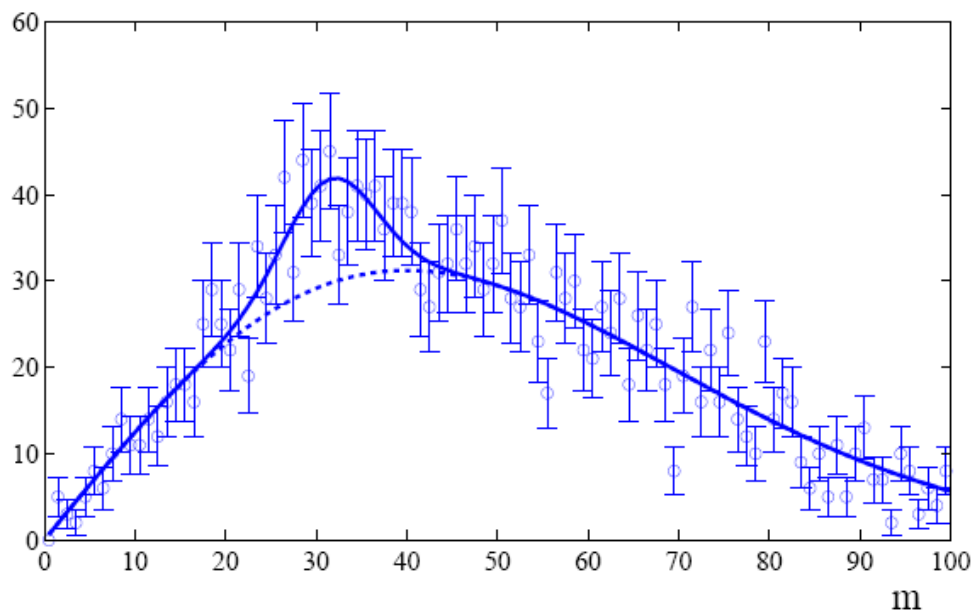
The interval's upper edge is higher than the limit from the one-sided test, and lower values of  $\mu$  may be excluded as well. A substantial downward fluctuation in the data gives a low (but nonzero) limit.

This means that when a value of  $\mu$  is excluded, it is because there is a probability  $\alpha$  for the data to fluctuate either high or low in a manner corresponding to less compatibility as measured by the likelihood ratio.

## The Look-Elsewhere Effect

Suppose a model for a mass distribution allows for a peak at a mass  $m$  with amplitude  $\mu$ .

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  $t_{\text{fix}}$  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

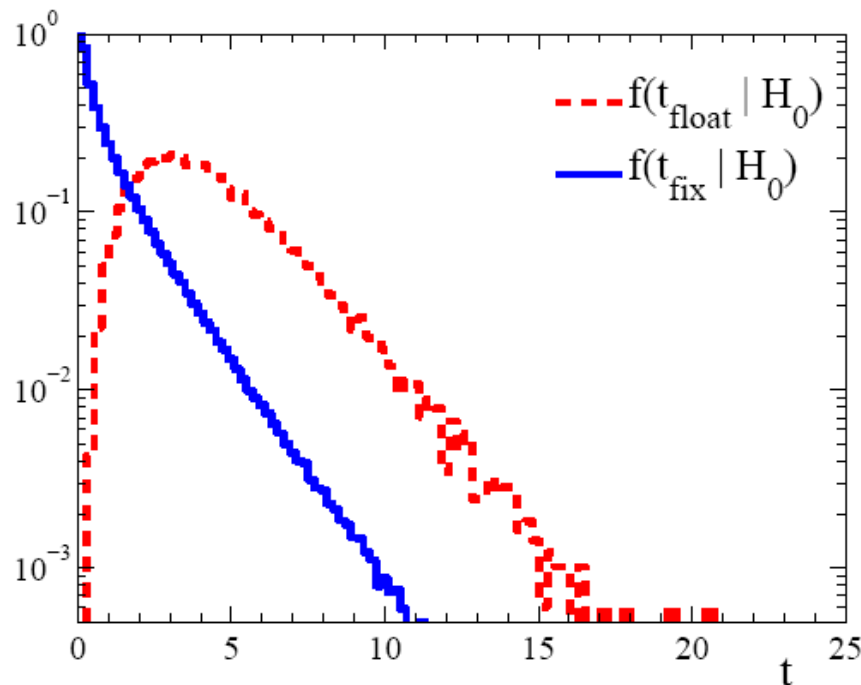
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \quad (\text{Note } m \text{ does not appear in the } \mu = 0 \text{ model.})$$

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

## Distributions of $t_{\text{fix}}$ , $t_{\text{float}}$

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

For  $t_{\text{float}}$  there are two adjustable parameters,  $\mu$  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  $t_{\text{float}}$  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

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

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).



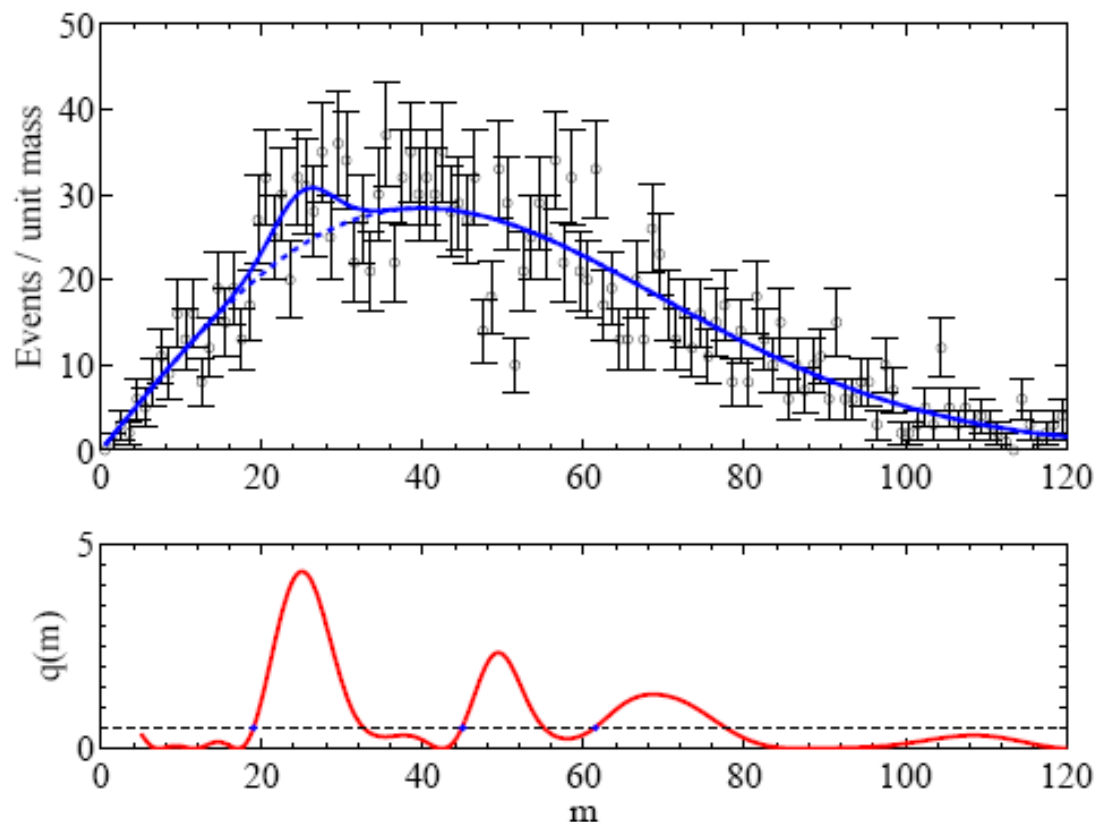
# Upcrossings of $-2\ln L$

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

$\langle N(c) \rangle$  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.

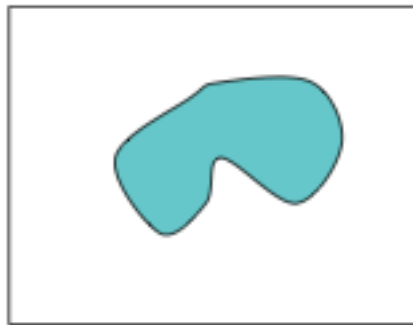


## Multidimensional look-elsewhere effect

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

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

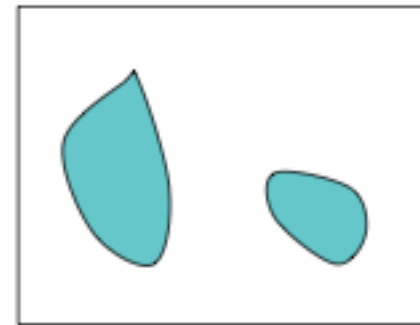
- Number of disconnected components minus number of 'holes'



$\varphi=1$



$\varphi=0$



$\varphi=2$

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

# 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 multiple 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 also problematic issue of testing many signal models (or parameter values) and thus excluding some for which one has little or no sensitivity.

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  $2.9 \times 10^{-7}$ , corresponding to a significance  $Z = \Phi^{-1}(1 - p) = 5$  (a  $5\sigma$  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  $3\sigma$  than  $5\sigma$ .