

# Introductory Statistics Course – Part II

<https://indico.cern.ch/event/735431/>



PHYSTAT – v

CERN

22-25 January 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

## Part I (Louis Lyons):

Combining results

Likelihoods

Bayes, Frequentism and F-C

MVA

Discovery issues

Upper Limits

## ⇒ Part II (GDC):

Nuisance parameters and systematic uncertainties

Tests from profile likelihood ratio

(More parameter estimation, Bayesian methods)

Experimental sensitivity

# Recap on tests

Consider a hypothesis  $H_0$  and relevant alternative  $H_1$ .

A **test** of  $H_0$  is defined by specifying a **critical region**  $w$  of the data space such that there is no more than some (small) probability  $\alpha$  (the size) assuming  $H_0$  is correct, to observe the data there, i.e.,

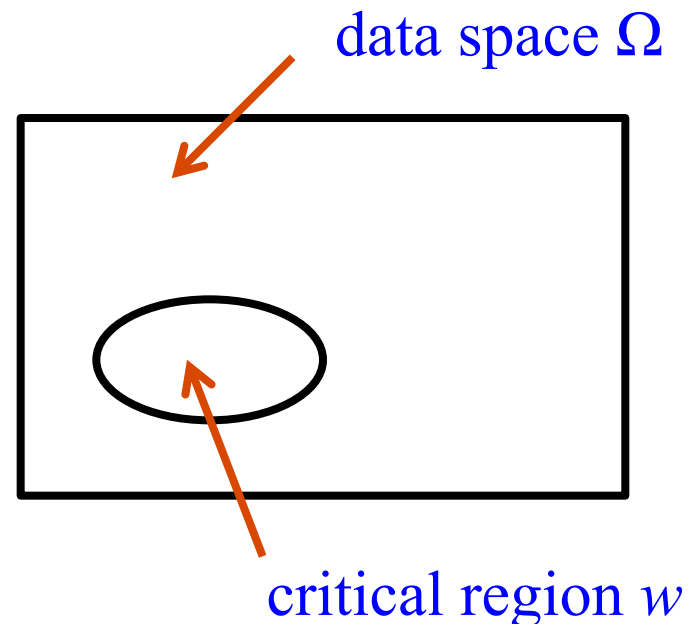
$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

Choose  $w$  to maximize the power with respect to alternative  $H_1$

$$M(H_1) = P(x \in w \mid H_1)$$

If  $x$  is observed in the critical region, reject  $H_0$ .



# Recap on testing significance / goodness-of-fit

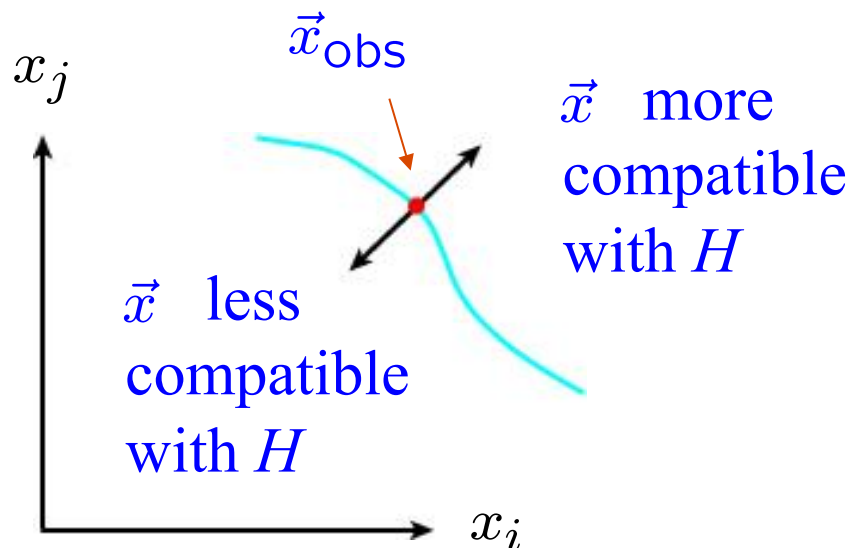
Suppose hypothesis  $H$  predicts pdf  $f(\vec{x}|H)$  for a set of observations  $\vec{x} = (x_1, \dots, x_n)$ .

We observe a single point in this space:  $\vec{x}_{\text{obs}}$

What can we say about the validity of  $H$  in light of the data?

Decide what part of the data space represents less compatibility with  $H$  than does the point  $\vec{x}_{\text{obs}}$ .

This region therefore has greater compatibility with some alternative  $H'$ .



# *p*-values

Express ‘goodness-of-fit’ by giving the *p*-value for *H*:

*p* = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.



This is not the probability that *H* is true!

In frequentist statistics we don’t talk about  $P(H)$  (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes’ theorem to obtain

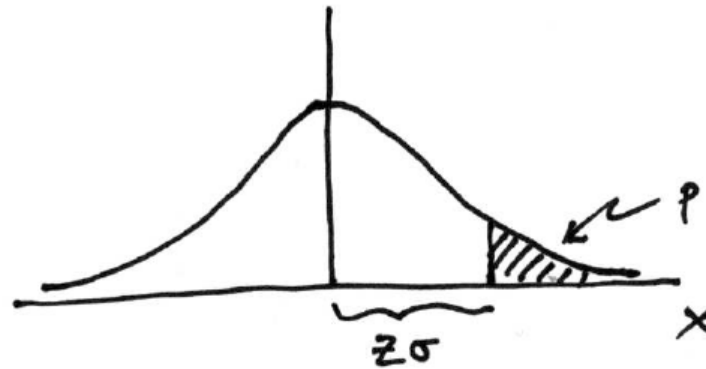
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) dH}$$

where  $\pi(H)$  is the prior probability for *H*.

For now stick with the frequentist approach; result is *p*-value, regrettably easy to misinterpret as  $P(H)$ .

# 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.



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

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

# Test statistics and $p$ -values

Consider a parameter  $\mu$  proportional to rate of signal process.

Often define a function of the data (test statistic)  $q_\mu$  that reflects level of agreement between the data and the hypothesized value  $\mu$ .

Usually define  $q_\mu$  so that higher values increasingly incompatibility with the data (more compatible with a relevant alternative).

We can define critical region of test of  $\mu$  by  $q_\mu \geq \text{const.}$ , or equivalently define the  $p$ -value of  $\mu$  as:

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

observed value of  $q_\mu$

pdf of  $q_\mu$  assuming  $\mu$

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

# Recap on confidence regions/intervals

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

The values that are not rejected constitute a *confidence region* (or interval) for  $\theta$  at confidence level  $CL = 1 - \alpha$ .

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

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

Usually use a  $p$ -value of  $\theta$  to define critical region of test as having  $p_\theta \leq \alpha$ .

The parameter values in the confidence region/interval have  $p$ -values of at least  $\alpha$ .

To find boundary of region/interval, set  $p_\theta = \alpha$  and solve for  $\theta$ .



# Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s)  $\theta = (\theta_1, \dots, \theta_n)$  using the ratio

$$\lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad 0 \leq \lambda(\theta) \leq 1$$

Lower  $\lambda(\theta)$  means worse agreement between data and hypothesized  $\theta$ . Equivalently, usually define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher  $t_\theta$  means worse agreement between  $\theta$  and the data.

$p$ -value of  $\theta$  therefore 
$$p_\theta = \int_{t_{\theta, \text{obs}}}^{\infty} f(t_\theta | \theta) dt_\theta$$

need pdf

# Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi_n^2$$

chi-square dist. with # d.o.f. =  
# of components in  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ .

Assuming this holds, the  $p$ -value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi_n^2}(t_{\boldsymbol{\theta}}) \quad \text{where} \quad F_{\chi_n^2}(t_{\boldsymbol{\theta}}) \equiv \int_0^{t_{\boldsymbol{\theta}}} f_{\chi_n^2}(t'_{\boldsymbol{\theta}}) dt'_{\boldsymbol{\theta}}$$

To find boundary of confidence region set  $p_{\boldsymbol{\theta}} = \alpha$  and solve for  $t_{\boldsymbol{\theta}}$ :

$$t_{\boldsymbol{\theta}} = F_{\chi_n^2}^{-1}(1 - \alpha) = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

## Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in  $\theta$  space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

For example, for  $1 - \alpha = 68.3\%$  and  $n = 1$  parameter,

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

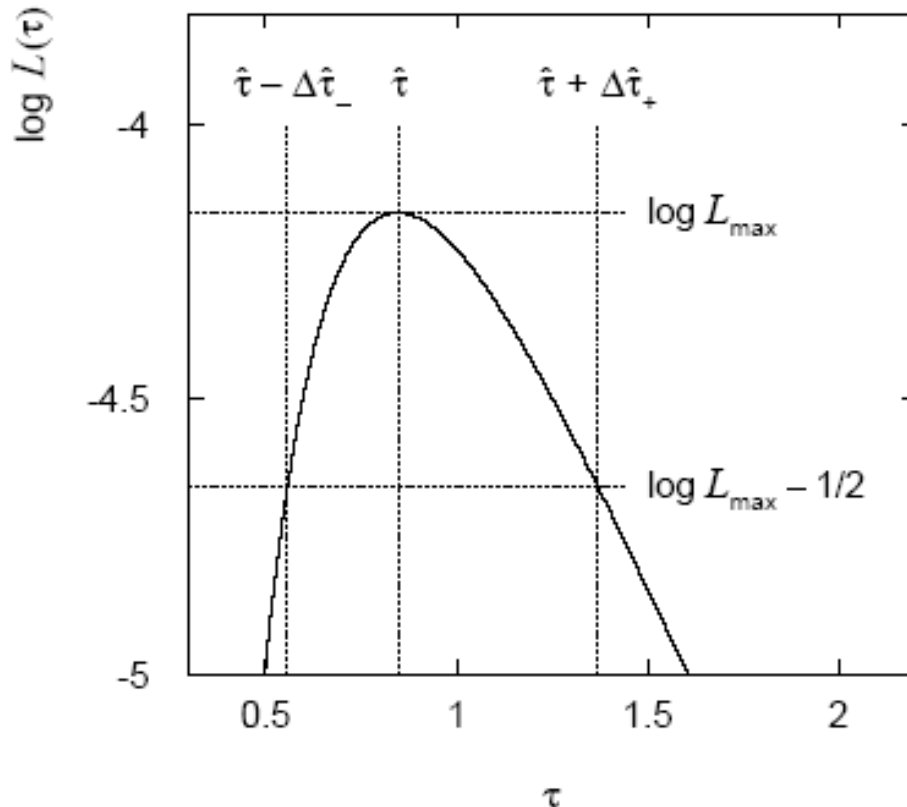
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$  is a 68.3% CL confidence interval.

# Example of interval from $\ln L$

For  $n = 1$  parameter,  $CL = 0.683$ ,  $Q_\alpha = 1$ .

Exponential example, now with only 5 events:



Parameter estimate and approximate 68.3% CL confidence interval:

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

# Multiparameter case

For increasing number of parameters,  $CL = 1 - \alpha$  decreases for confidence region determined by a given

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

$Q_\alpha$	$1 - \alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

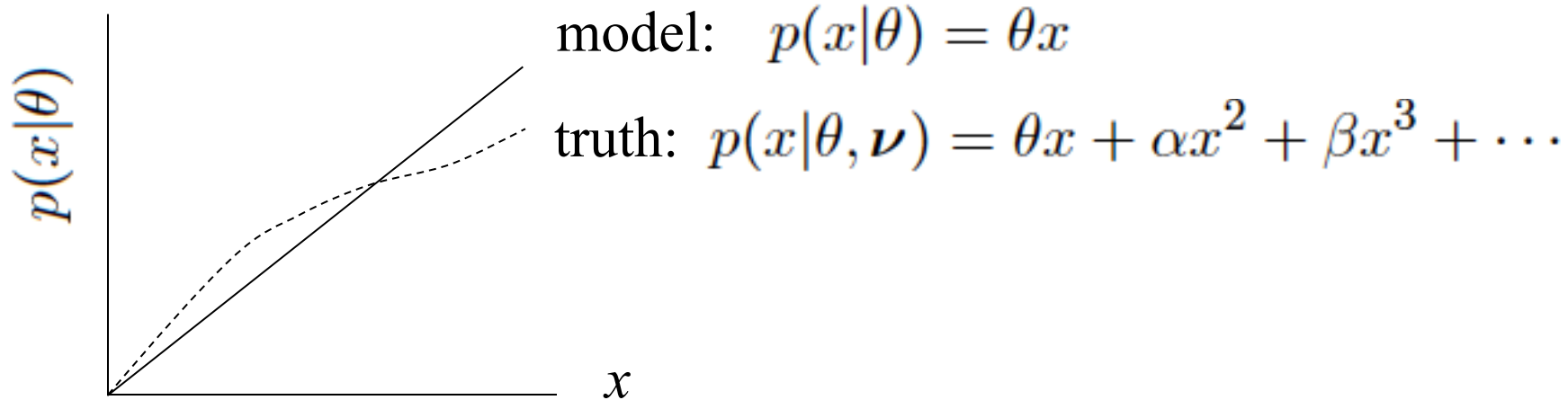
## Multiparameter case (cont.)

Equivalently,  $Q_\alpha$  increases with  $n$  for a given  $\text{CL} = 1 - \alpha$ .

$1 - \alpha$	$\bar{Q}_\alpha$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

# 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).

# $p$ -values in cases with nuisance parameters

Suppose we have a statistic  $q_\theta$  that we use to test a hypothesized value of a parameter  $\theta$ , such that the  $p$ -value of  $\theta$  is

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

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

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

→ “exact” confidence interval

But in general for finite data samples this is not true; one may be unable to reject some  $\theta$  values if all values of  $\nu$  must be considered (resulting interval for  $\theta$  “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  $\theta$  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  $\theta$ :

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

“double hat” notation means profiled value, i.e., parameter that maximizes likelihood for the given  $\theta$ .

The resulting confidence interval will have the correct coverage for the points  $(\theta, \hat{\hat{v}}(\theta))$ .

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).

# 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  $\boldsymbol{\nu}$** .

# 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:

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

maximizes  $L$  for specified  $\mu$

maximize  $L$

Define critical region of test of  $\mu$  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

Suppose relevant alternative to background-only ( $\mu = 0$ ) is  $\mu \geq 0$ .

So take critical region for test of  $\mu = 0$  corresponding to high  $q_0$  and  $\hat{\mu} > 0$  (data characteristic for  $\mu \geq 0$ ).

That is, to test background-only hypothesis define statistic

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

i.e. here only large (positive) observed signal strength is 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:

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

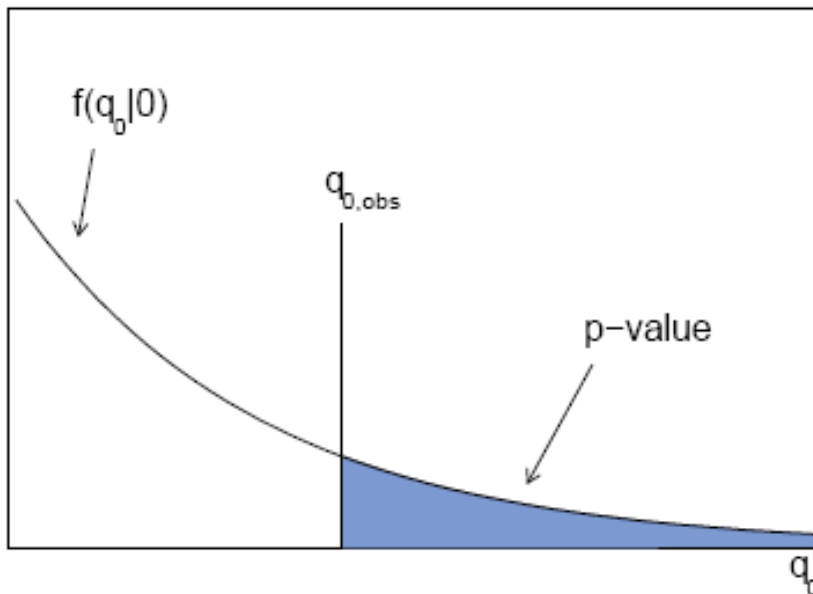
In large sample limit,  $f(q_0|0)$  independent of nuisance parameters;  $f(q_0|\mu')$  depends on nuisance parameters through  $\sigma$ .

# *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

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

# Monte Carlo test of asymptotic formula

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

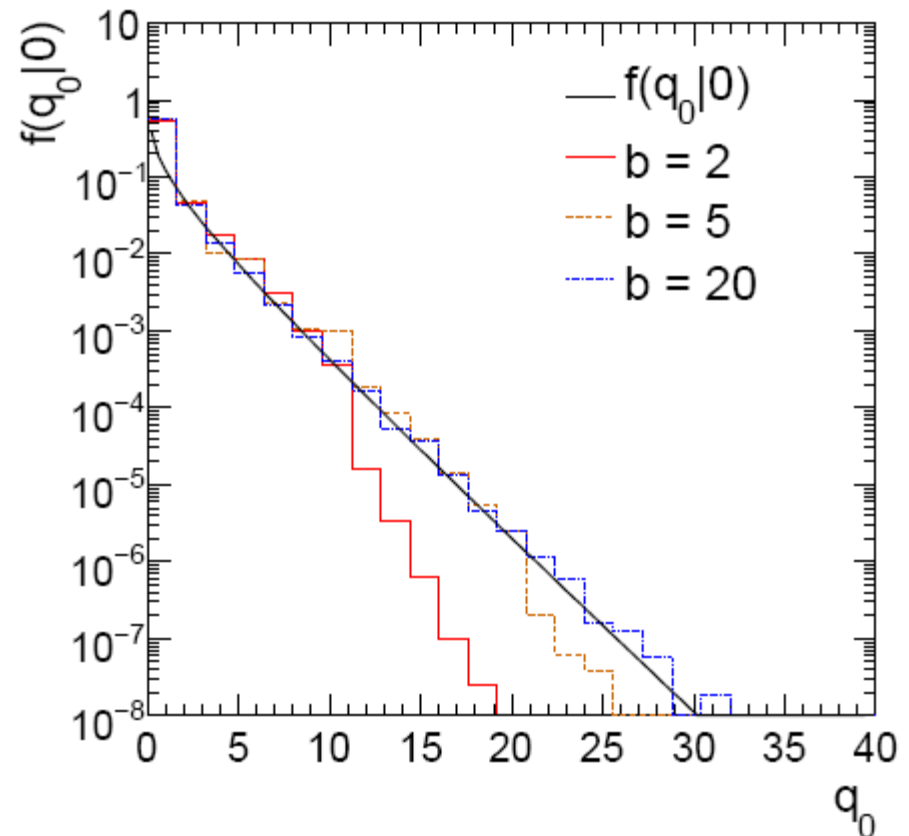
$$m \sim \text{Poisson}(\tau b)$$

$\mu$  = param. of interest

$b$  = nuisance parameter

Here take  $s$  known,  $\tau = 1$ .

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



## Test statistic for upper limits

For purposes of setting an upper limit on  $\mu$  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

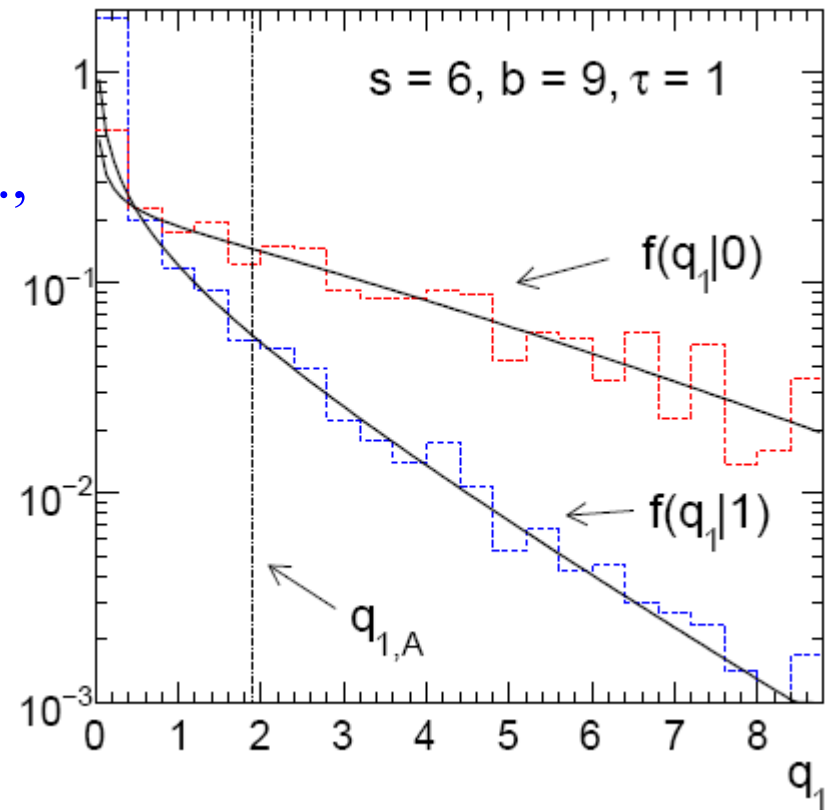
Consider again  $n \sim \text{Poisson}(\mu s + b)$ ,  $m \sim \text{Poisson}(\tau b)$   
 Use  $q_\mu$  to find  $p$ -value of hypothesized  $\mu$  values.

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  $s = 6$ ,  $b = 9$ .



# Unified intervals from likelihood ratio

Suppose relevant alternative to tested value of  $\mu$  could be higher or lower. 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  $\mu$ .

## 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

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

# Unified (Feldman-Cousins) intervals

If negative  $\mu$  not allowed, can use

$$\tilde{t}_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(0, \hat{\theta}(0))} & \hat{\mu} < 0, \\ -2 \ln \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})} & \hat{\mu} \geq 0. \end{cases}$$

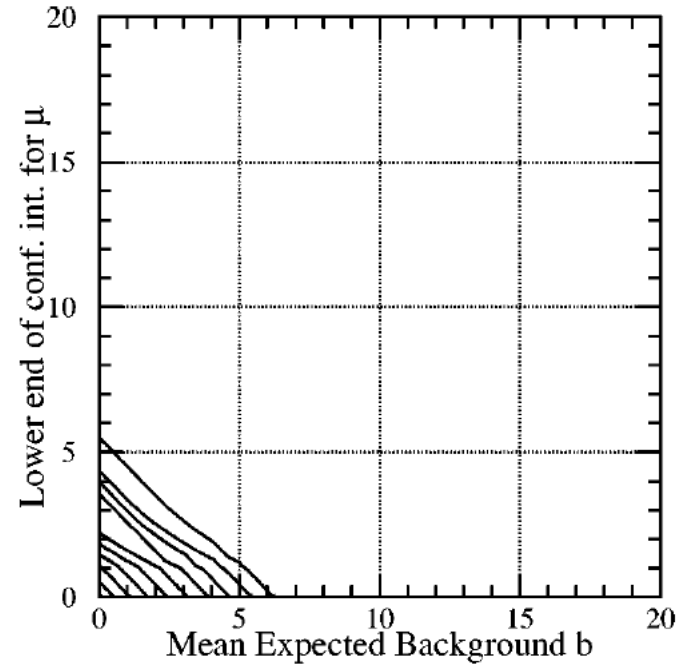
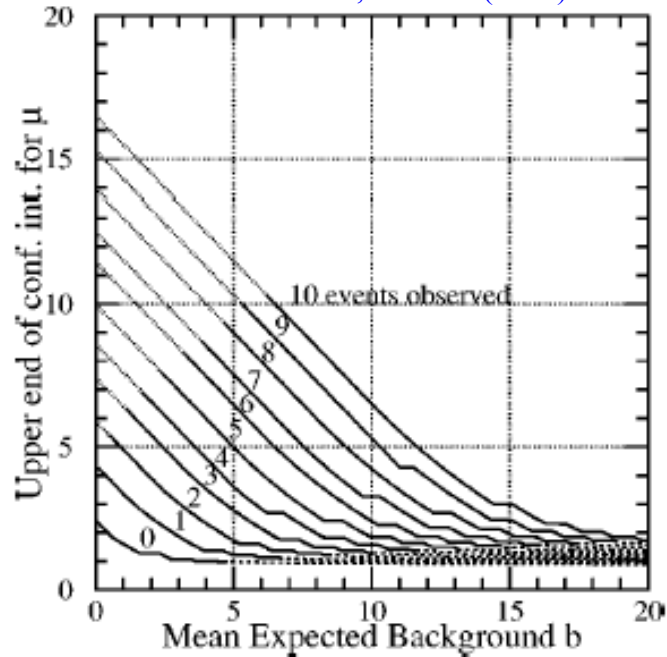
This is essentially the statistic used for Feldman-Cousins intervals (here also treats nuisance parameters).

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

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

# 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  $b$ .



# 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:

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

# 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),

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

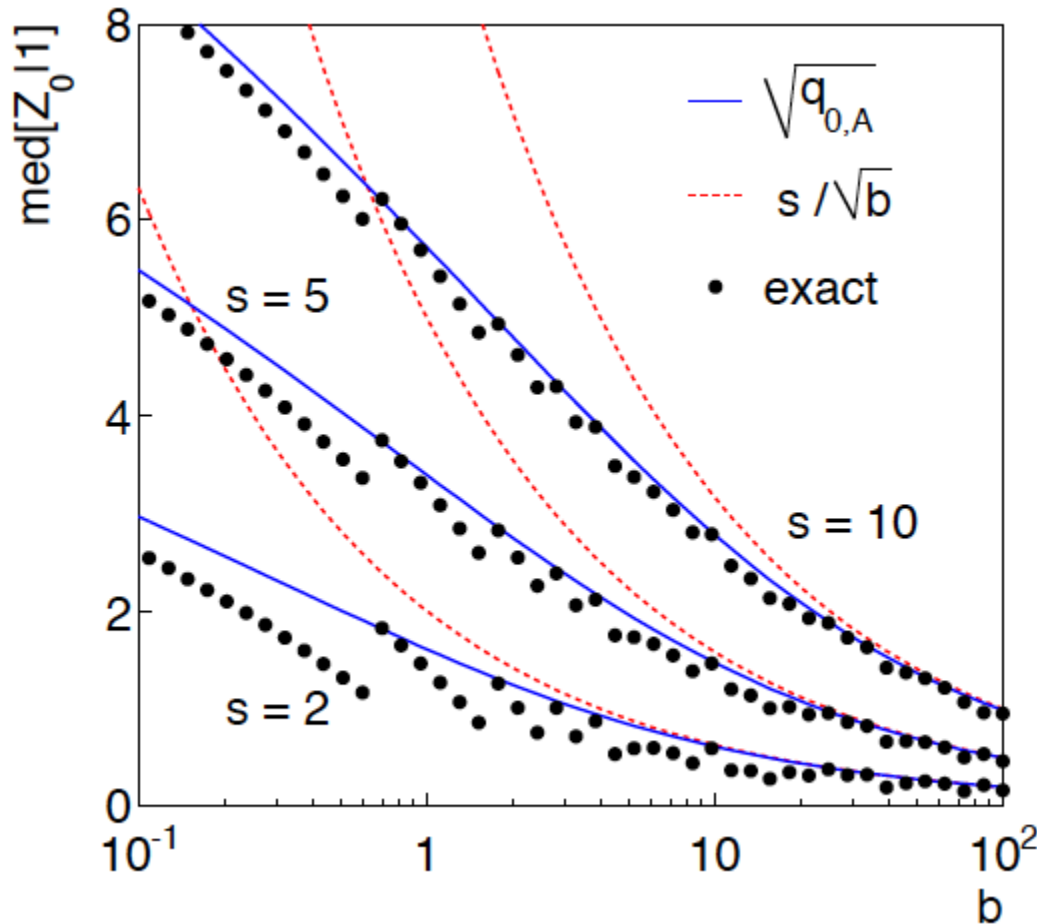
To find  $\text{median}[Z|s]$ , let  $n \rightarrow s + b$  (i.e., the Asimov data set):

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

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  $\tau$ :

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

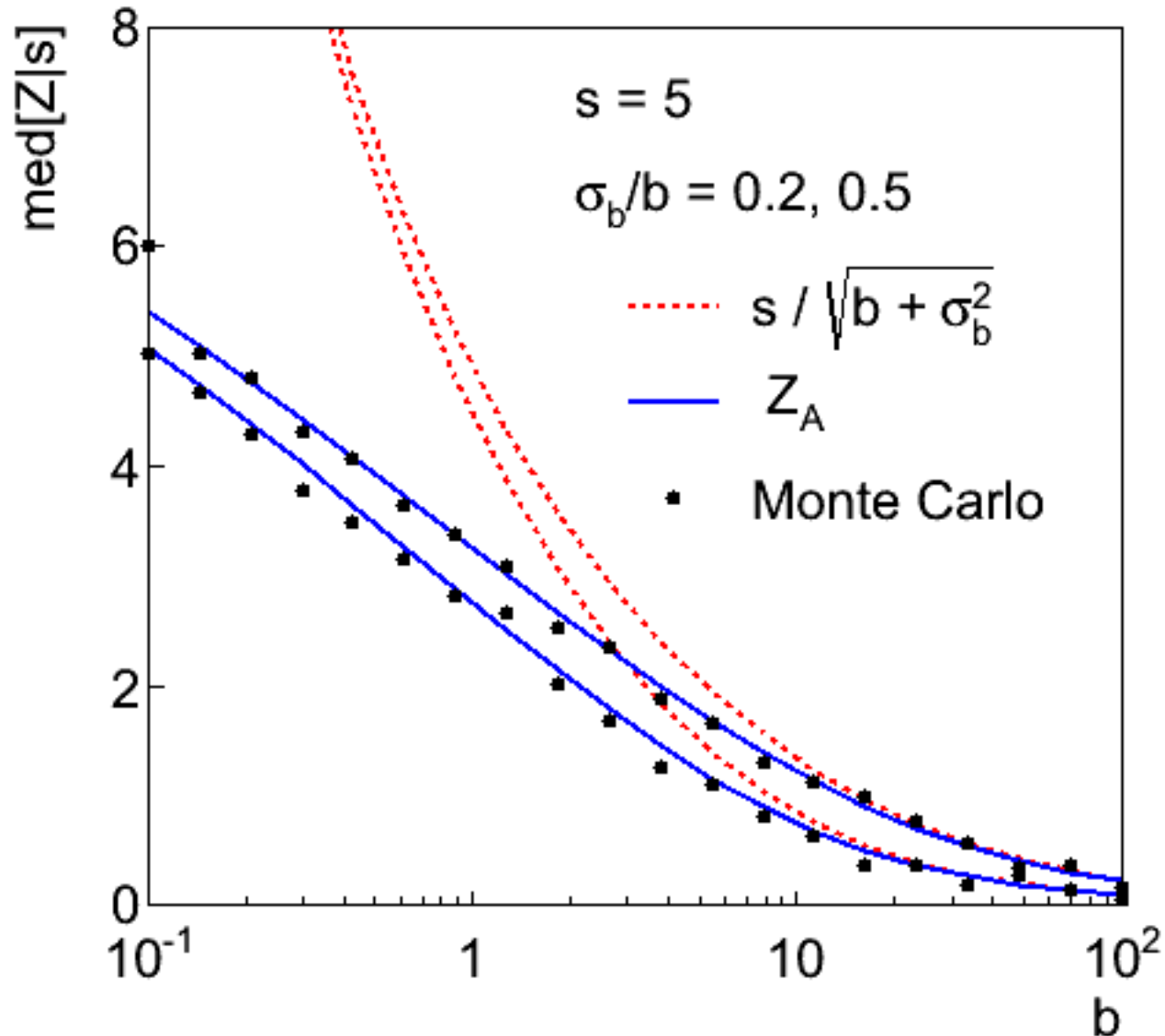
# Limiting cases

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

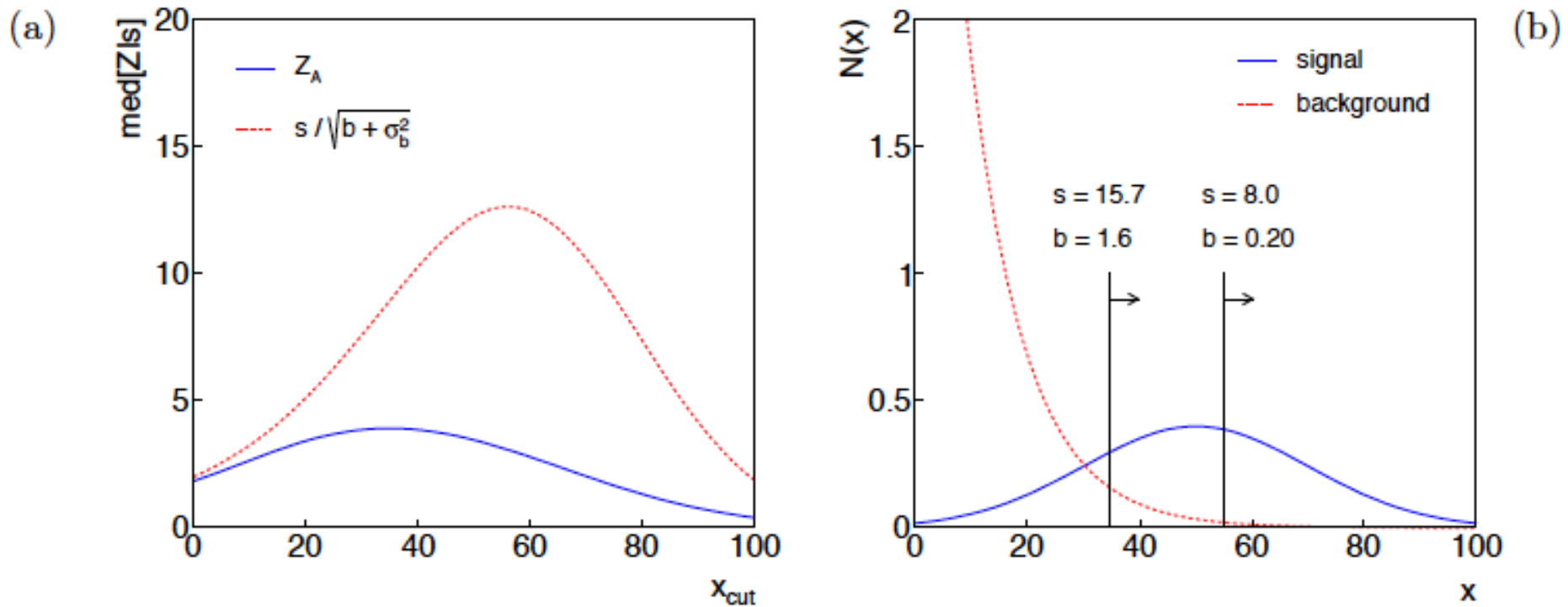


Figure 1: (a) The expected significance as a function of the cut value  $x_{\text{cut}}$ ; (b) the distributions of signal and background with the optimal cut value indicated.

# 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).

# Finally

Here only a brief introduction to selected topics

Nuisance parameters and systematic uncertainties

Tests from profile likelihood ratio

More parameter estimation, Bayesian methods

Experimental sensitivity

No time today for many important issues

Limits in case of low sensitivity

Bayesian approach to discovery (Bayes factors)

Unfolding, etc., etc.

Focus should be on writing down the likelihood, e.g.,  $P(x|\theta)$ , and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).



# Extra slides

# Example: fitting a straight line

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

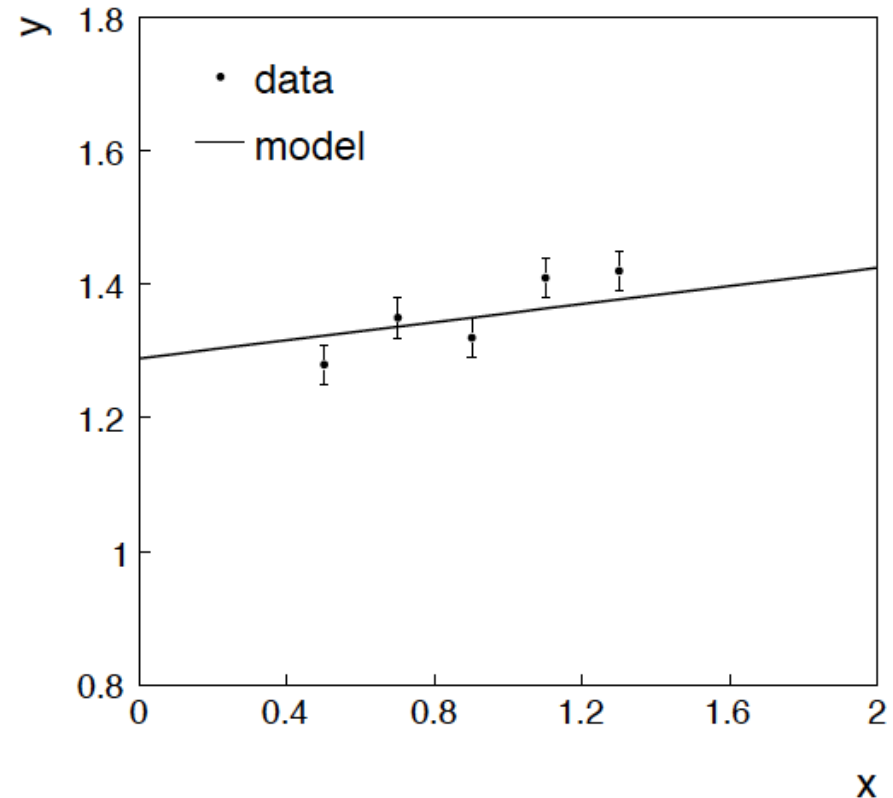
Model:  $y_i$  independent and all follow  $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

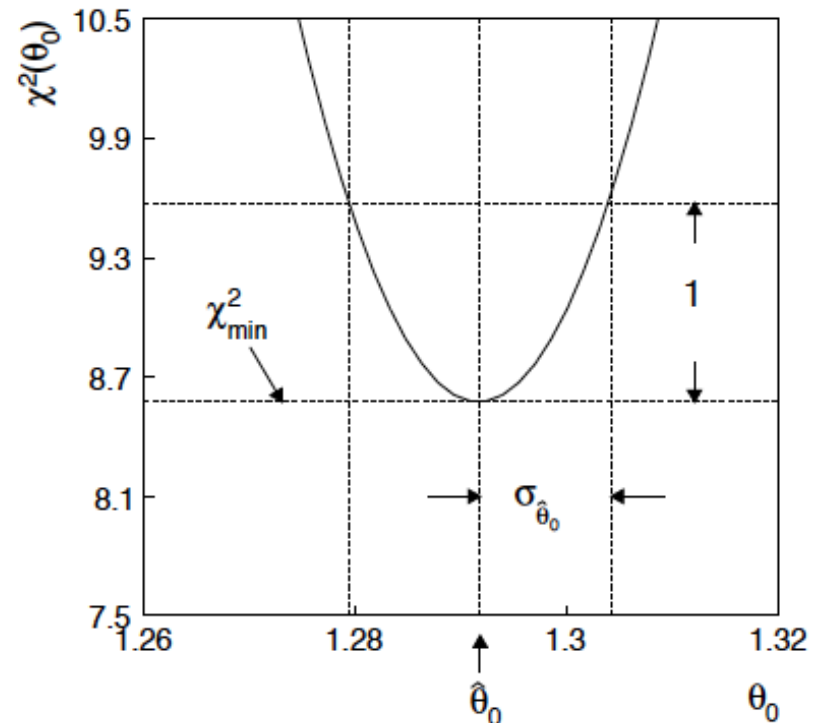
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

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 $\theta_1$

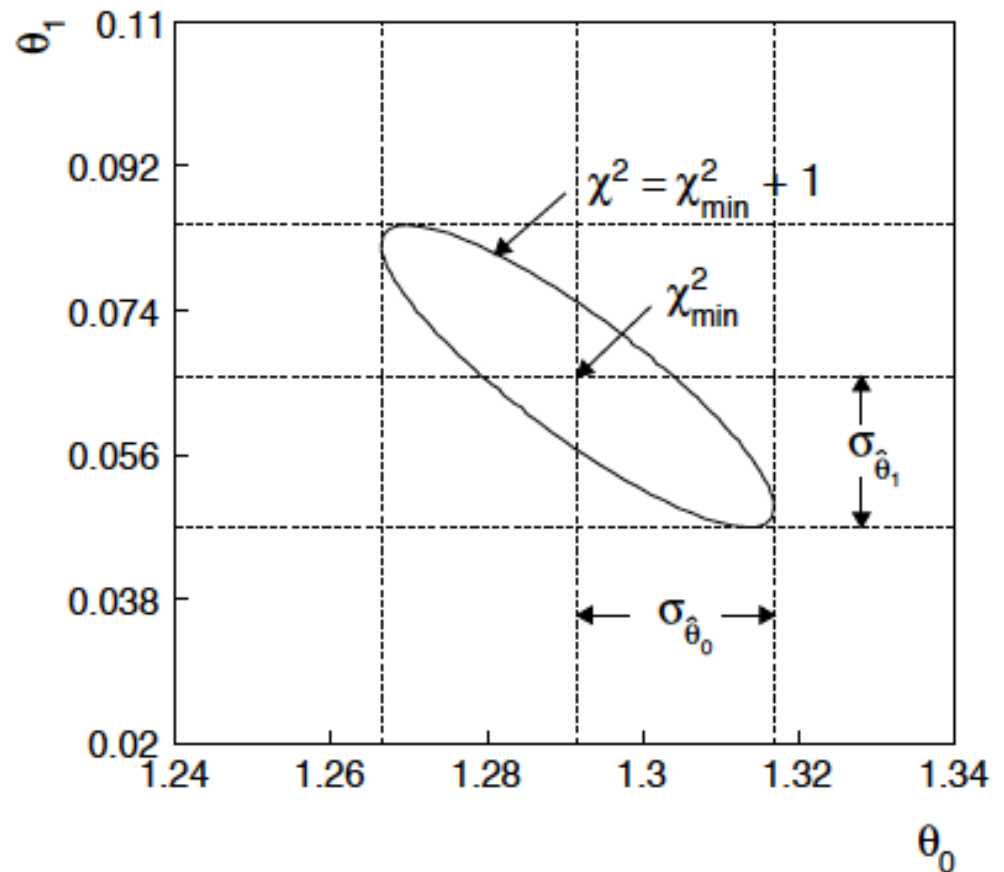
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

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.

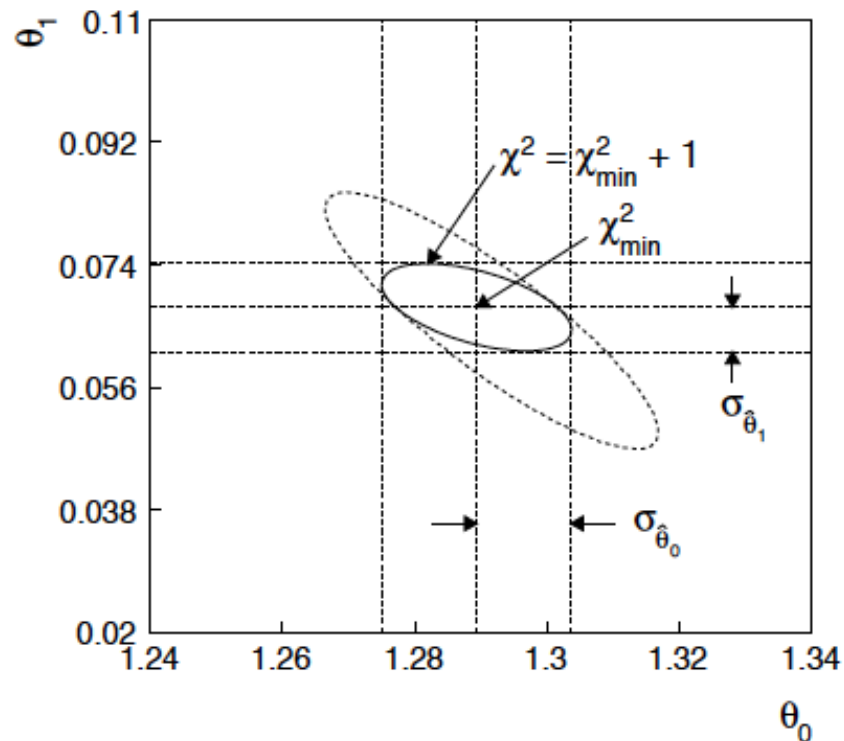


If we have a measurement  $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

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

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

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
∝
likelihood
×
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:  $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

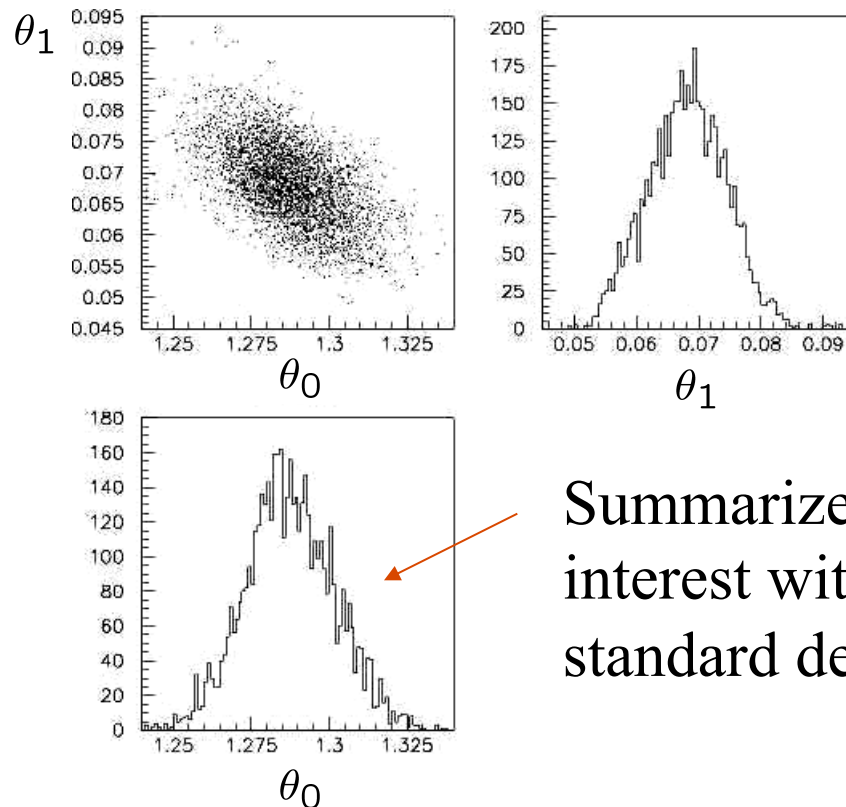
Test ratio is (*Metropolis-Hastings*):  $\alpha = \min \left[ 1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

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.

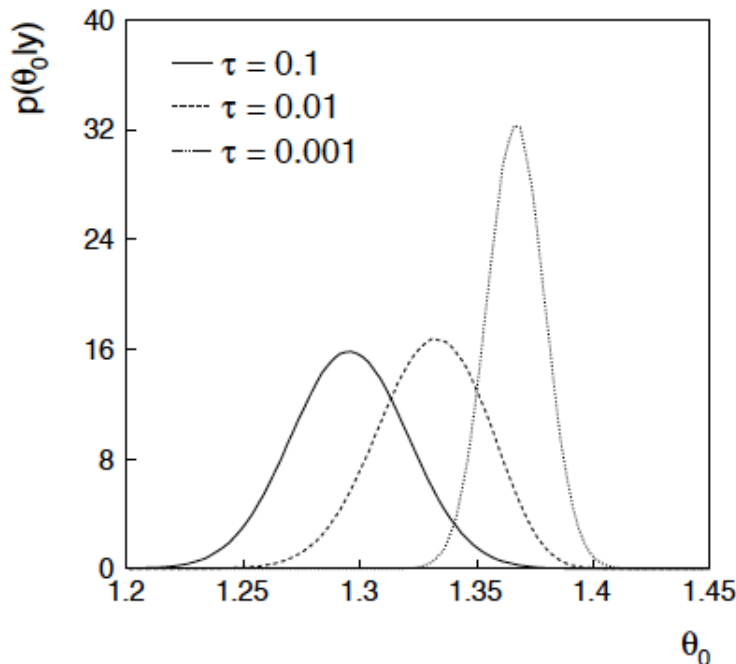
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

# 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.

# 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.

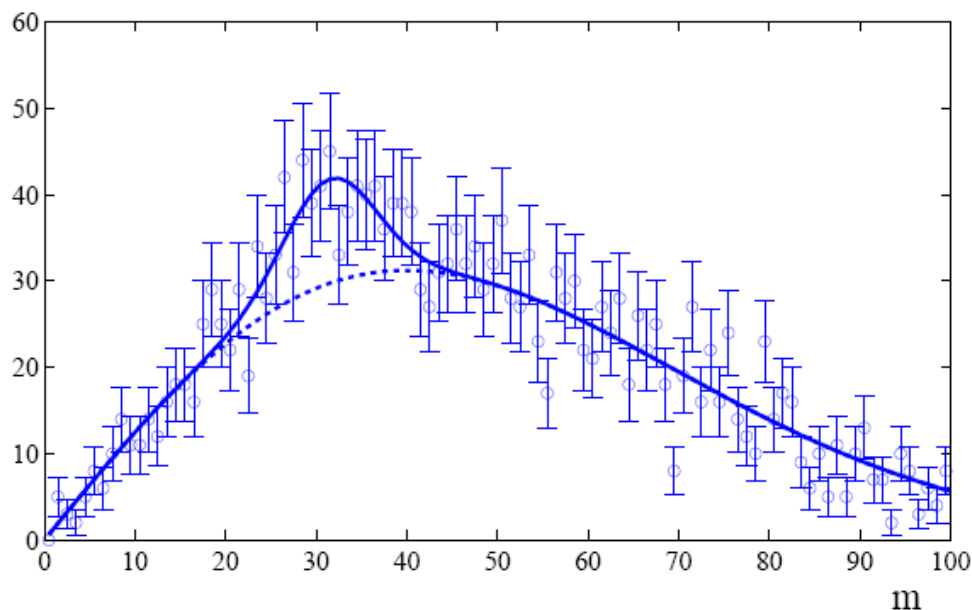
Providing LEE is dealt with, that threshold is probably closer to  $3\sigma$  than  $5\sigma$ .



## 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

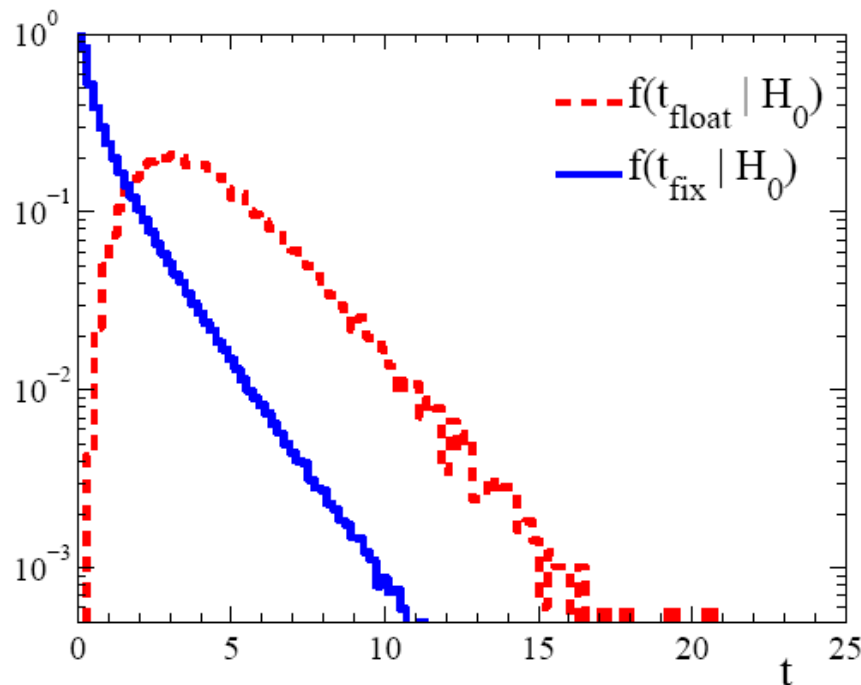
$$t_{\text{float}} = -2 \ln \frac{L(0)}{L(\hat{\mu}, \hat{m})} \quad (\text{Note } m \text{ does not appear in the } \mu = 0 \text{ model.})$$

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

## 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).

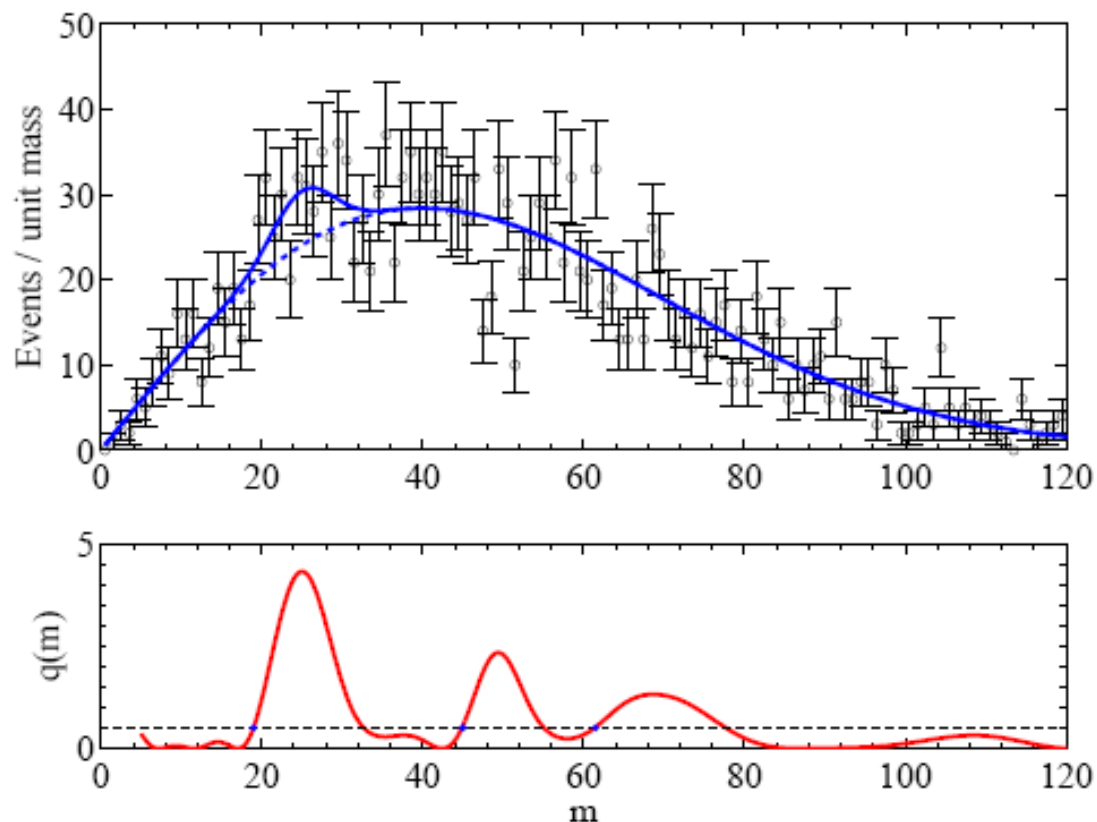
## Upcrossings of $-2\ln L$

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.

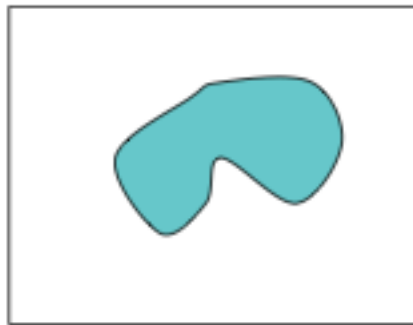


## Multidimensional look-elsewhere effect

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

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

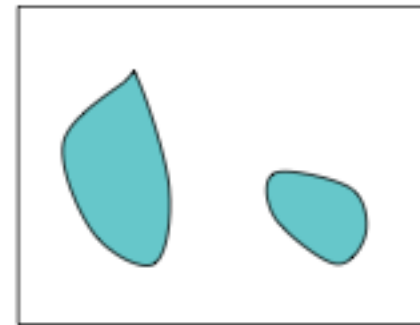
- Number of disconnected components minus number of 'holes'



$\varphi=1$



$\varphi=0$



$\varphi=2$

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