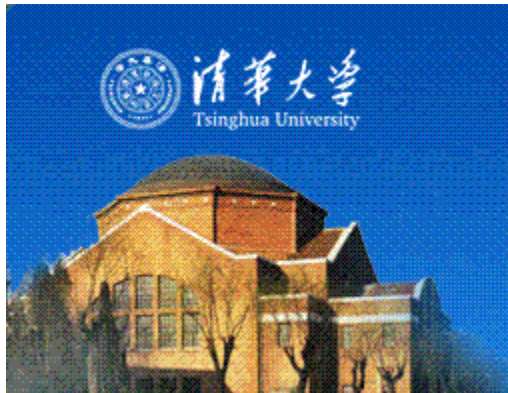


Statistical Methods in Particle Physics

Day 4: Discovery and limits



清华大学高能物理研究中心
2010年4月12—16日



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

Outline of lectures

Day #1: Introduction

Review of probability and Monte Carlo

Review of statistics: parameter estimation

Day #2: Multivariate methods (I)

Event selection as a statistical test

Cut-based, linear discriminant, neural networks

Day #3: Multivariate methods (II)

More multivariate classifiers: BDT, SVM ,...

→ Day #4: **Significance tests for discovery and limits**

Including systematics using profile likelihood

Day #5: Bayesian methods

Bayesian parameter estimation and model selection

Day #4: outline

Significance tests, p -values

- Significance test for discovery

- Setting limits

Including systematic uncertainties (frequentist)

- Profile likelihood function

- Examples of including nuisance parameters

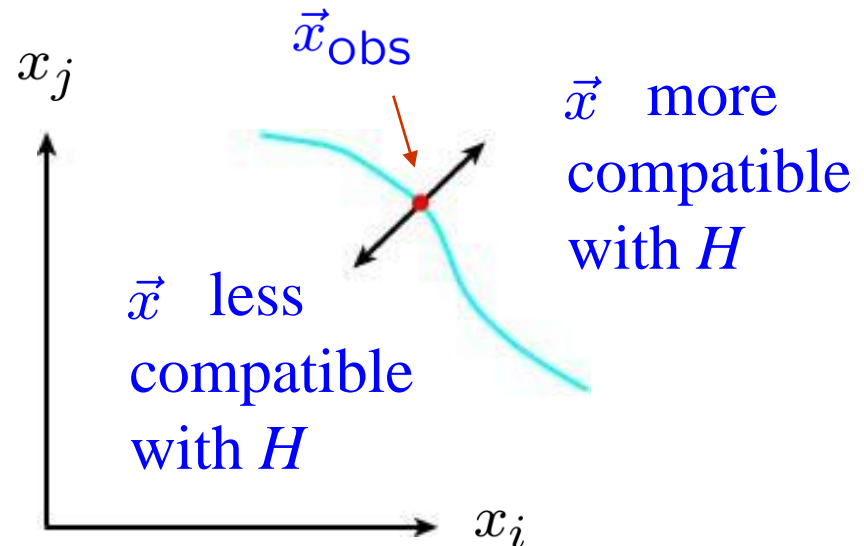
Testing significance / goodness-of-fit

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}_{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}_{obs} .
(Not unique!)



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 $P(H)$.

p-value example: testing whether a coin is ‘fair’

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N - n)!} p^n (1 - p)^{N-n}$$

Hypothesis H : the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) ‘by chance’, under the assumption of H .

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)

n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b , then $n = n_s + n_b$ is also Poisson, mean = $s + b$:

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

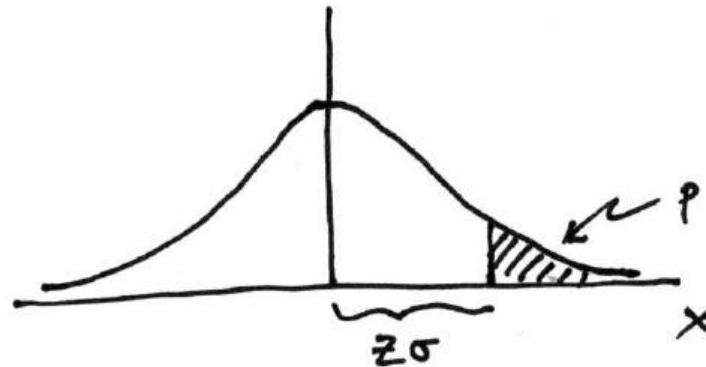
Suppose $b = 0.5$, and we observe $n_{\text{obs}} = 5$. Should we claim evidence for a new discovery?

Give p -value for hypothesis $s = 0$:

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$

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.



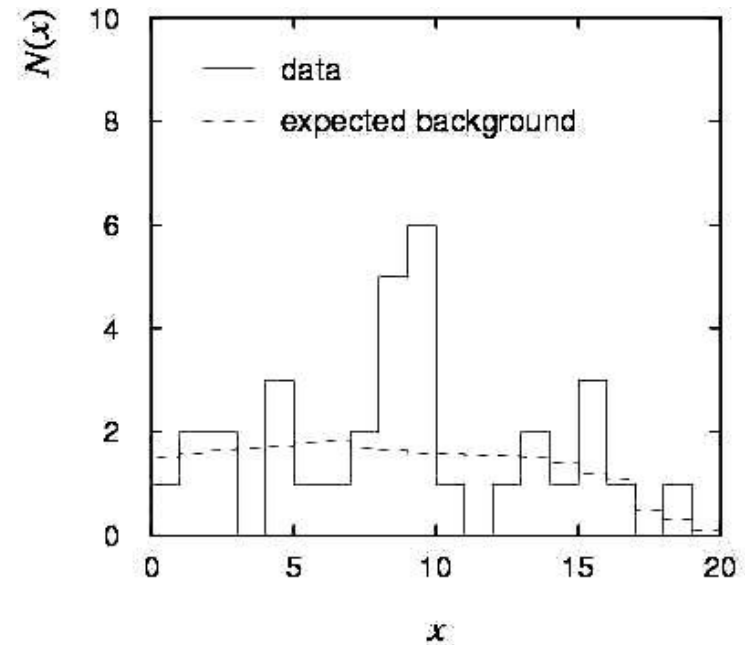
$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \text{TMath::Prob}$$

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

The significance of a peak

Suppose we measure a value x for each event and find:

Each bin (observed) is a Poisson r.v., means are given by dashed lines.



In the two bins with the peak, 11 entries found with $b = 3.2$.
The p -value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins \times distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

<u>phenomenon</u>	<u>reasonable p-value for discovery</u>
D ⁰ D ⁰ mixing	~0.05
Higgs	~ 10 ⁻⁷ (?)
Life on Mars	~10 ⁻¹⁰
Astrology	~10 ⁻²⁰

Setting limits: Poisson data with background

Count n events, e.g., in fixed time or integrated luminosity.

s = expected number of signal events

b = expected number of background events

$$n \sim \text{Poisson}(s+b): \quad P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

Suppose the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{\text{obs}} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s , taking into consideration any uncertainty in b .

Setting limits

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

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

If n is observed in the critical region, reject the value s .

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

set of s 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 s with probability $\geq 1 - \gamma$.

Frequentist upper limit for Poisson parameter

First suppose that the expected background b is known.

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a **confidence level** of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \rightarrow s_{\text{up}} = -\ln \gamma \approx 3.00$$

$[0, s_{\text{up}}]$ is an example of a **confidence interval**. It is designed to include the true value of s with probability at least $1-\gamma$ for any s .

Calculating Poisson parameter limits

Analogous procedure for lower limit s_{lo} .

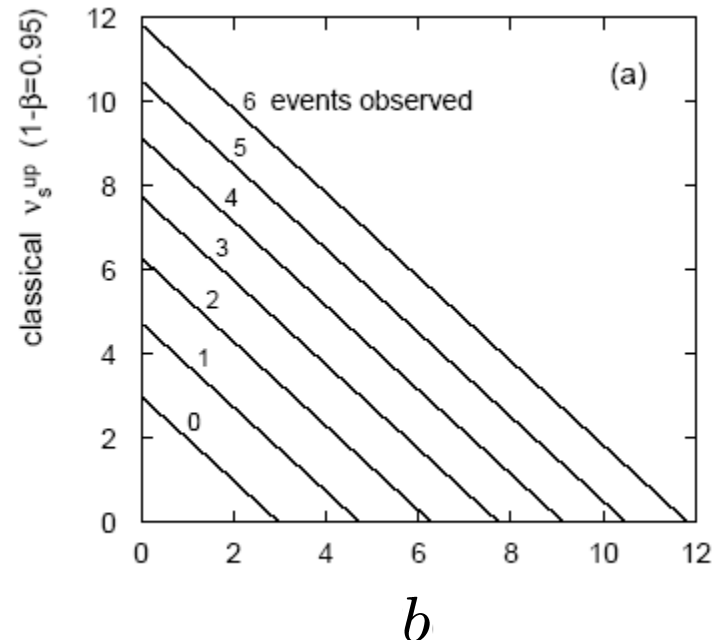
To solve for s_{lo} , s_{up} , can exploit relation to χ^2 distribution:

$$s_{lo} = \frac{1}{2} F_{\chi^2}^{-1}(\alpha; 2n) - b$$

Quantile of χ^2 distribution

$$s_{up} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \beta; 2(n + 1)) - b$$

For low fluctuation of n this can give negative result for s_{up} ; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose $CL = 0.9$, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary, cf. m_ν estimated using $E^2 - p^2$.

Expected limit for on s if $s = 0$

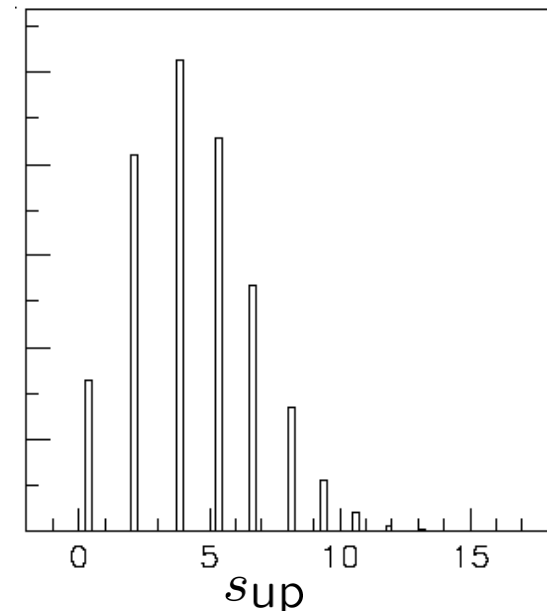
Physicist: I should have used $CL = 0.95$ — then $s_{up} = 0.496$

Even better: for $CL = 0.917923$ we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



Prototype LHC 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:

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

(N.B. here m = number of counts, not mass!)

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

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^N \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^M \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

Profile likelihood ratio

To test hypothesized value of μ , construct **profile likelihood ratio**:

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

← Maximized L for given μ

← Maximized L

Equivalently use $q_\mu = -2 \ln \lambda(\mu)$:

data agree well with hypothesized $\mu \rightarrow q_\mu$ small

data disagree with hypothesized $\mu \rightarrow q_\mu$ large

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

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p -value for an observed $q_{0,\text{obs}}$ is

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

Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_{\mu} = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

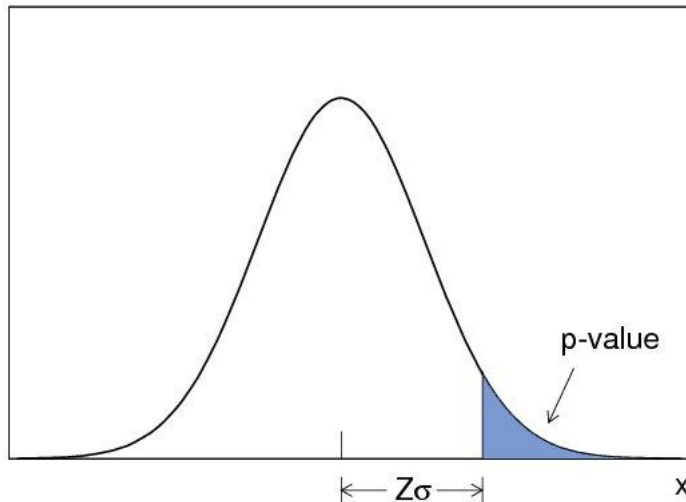
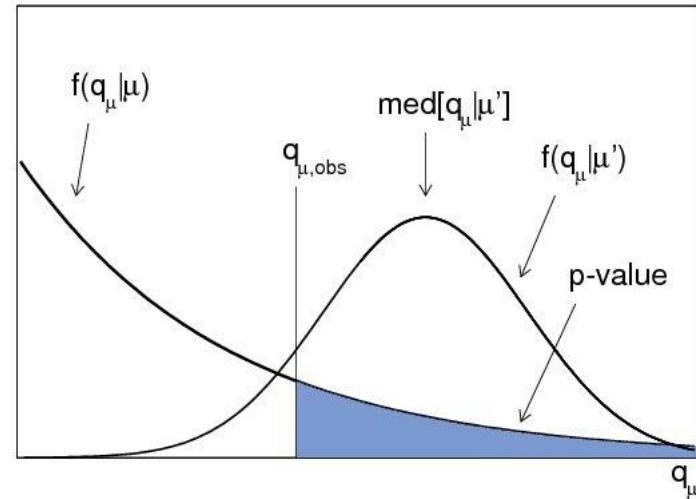
Note for purposes of setting an upper limit, one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ .

p -value of hypothesized μ is (similarly to the case of discovery)

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

p -value / significance of hypothesized μ

Test hypothesized μ by giving p -value, probability to see data with \leq compatibility with μ compared to data observed:



Equivalently use **significance**, Z , defined as equivalent number of sigmas for a Gaussian fluctuation in one direction:

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

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.,

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

Distribution of q_0

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

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

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

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

Distribution of q_μ

Similar results for q_μ

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

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

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

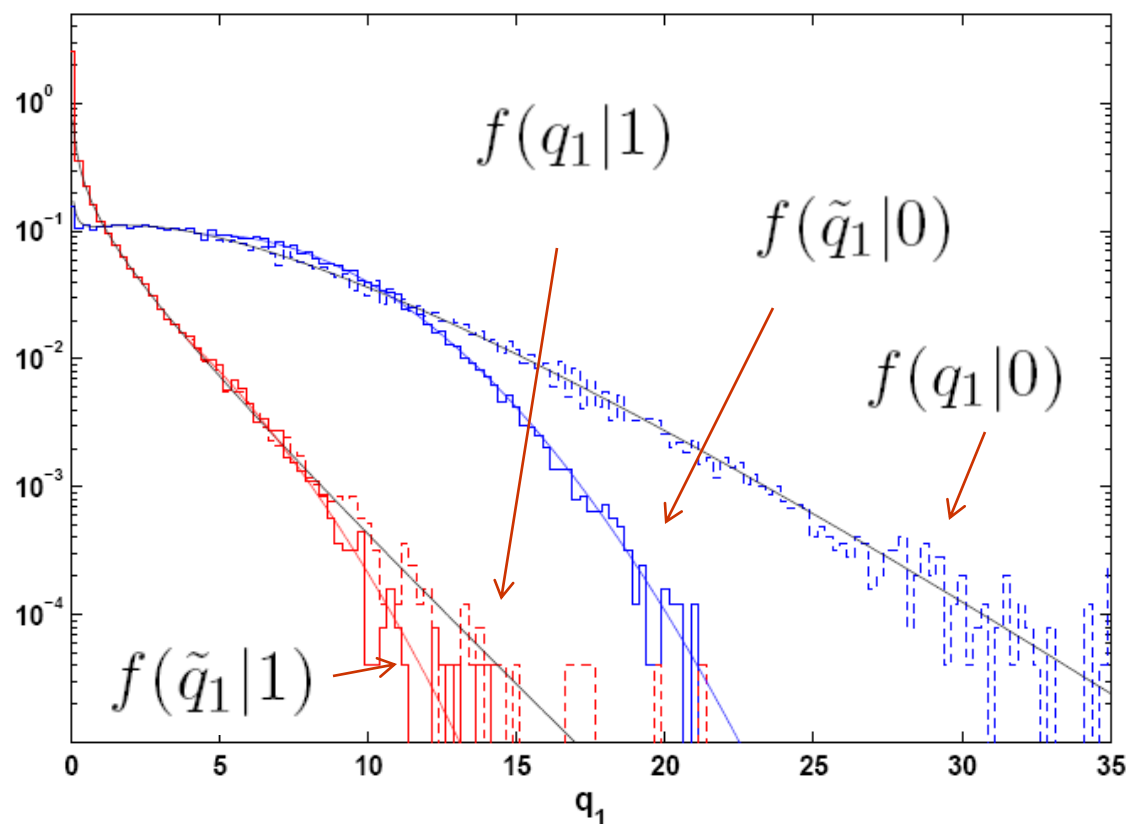
$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

E.g. if $p_\mu < 0.05$, μ is excluded at 95% CL.

An example

$$n \sim \text{Poisson}(\mu s + b) \quad s = 50, b = 100, \tau = 1$$

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



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 $\text{Prob}(x > x_{\text{obs}} | s = 0)$,:

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

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 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

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

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

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

To find $\text{median}[Z_0|s+b]$, let $n \rightarrow s + b$,

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

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

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):

$$H \rightarrow \gamma\gamma$$

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

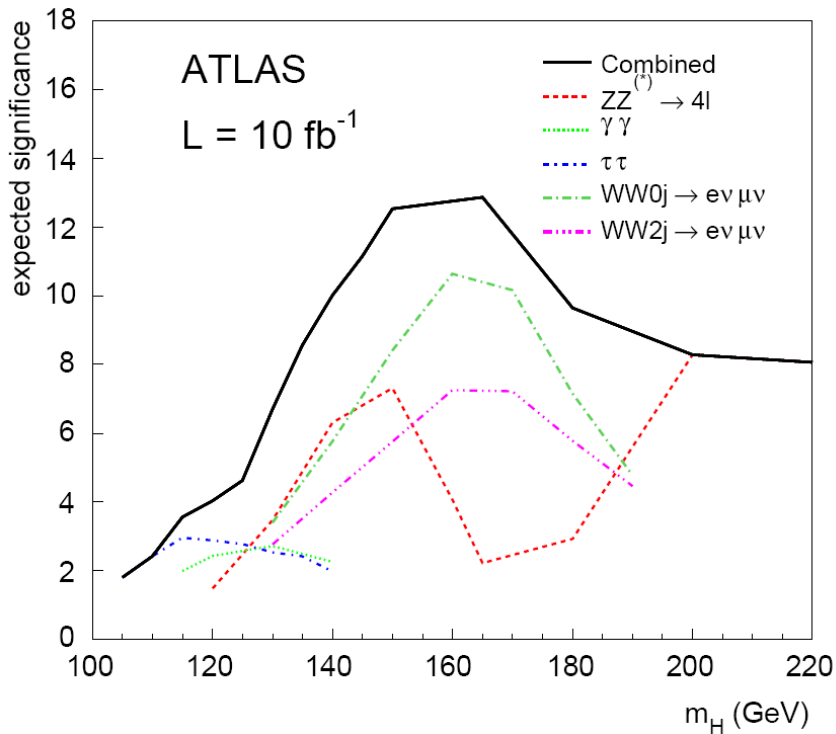
$$H \rightarrow ZZ^{(*)} \rightarrow 4l \quad (l = e, \mu)$$

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

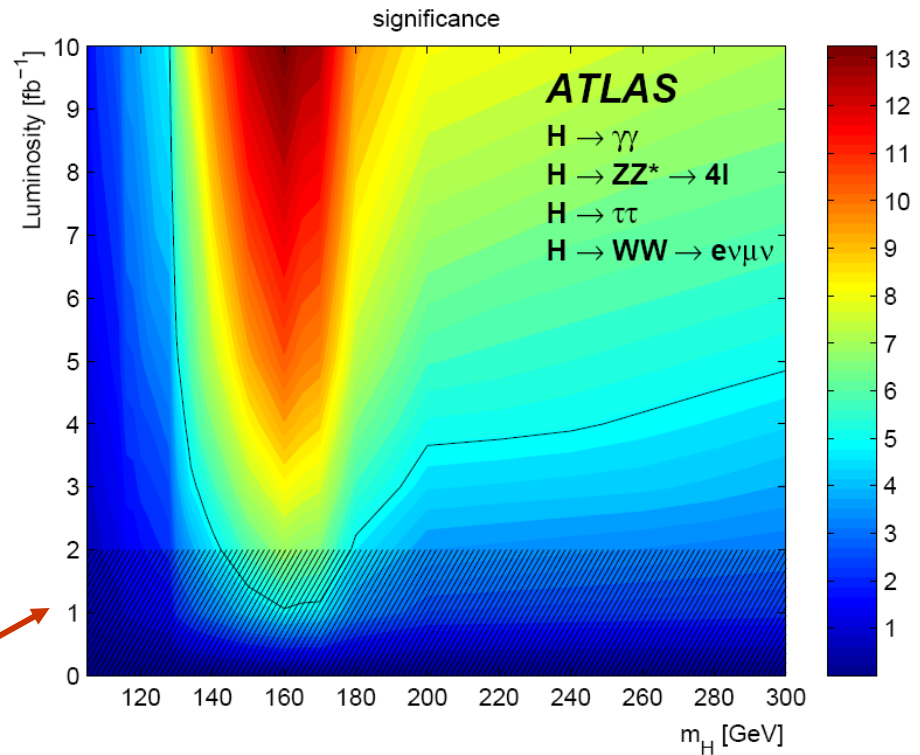
Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

Combined discovery significance



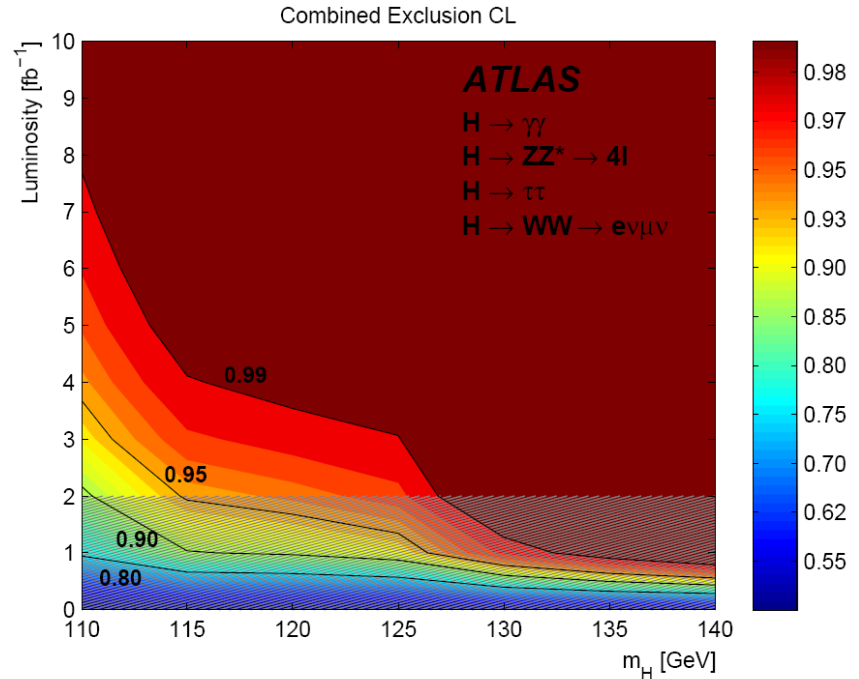
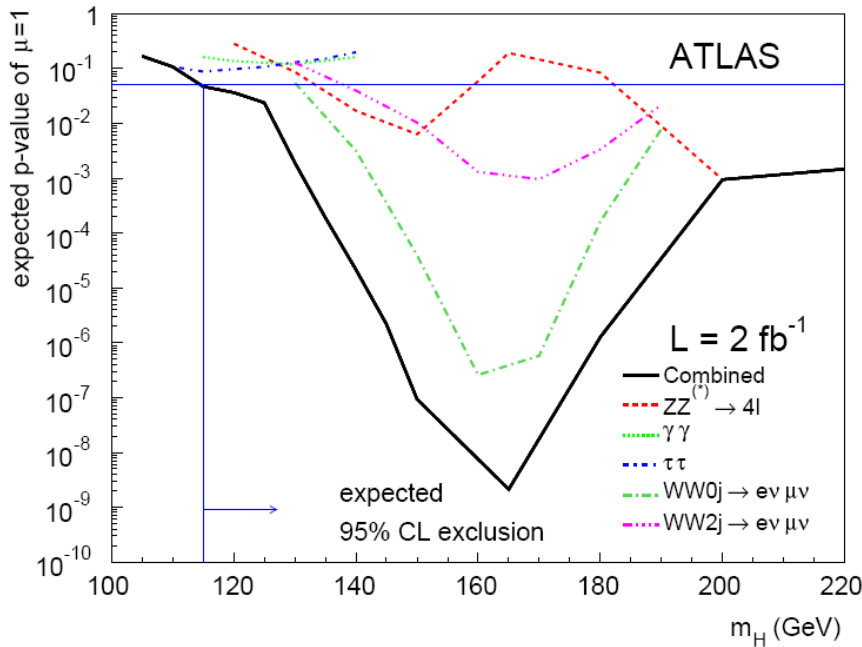
Discovery significance
(in colour) vs. L, m_H :



Approximations used here not always accurate for $L < 2 \text{ fb}^{-1}$ but in most cases conservative.

Combined 95% CL exclusion limits

$1 - p$ -value of m_H
(in colour) vs. L, m_H :



Summary on limits

Different sorts of limits answer different questions.

A frequentist confidence interval does not (necessarily) answer, “What do we believe the parameter’s value is?”

Look at sensitivity, e.g., $E[s_{\text{up}} | s = 0]$; consider also:

need for consensus/conventions;
convenience and ability to combine results, ...

For any result, consumer will compute (mentally or otherwise):

$$p(\theta|\text{result}) \propto L(\text{result}|\theta)\pi(\theta)$$

Need likelihood (or summary thereof).

consumer’s prior



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.

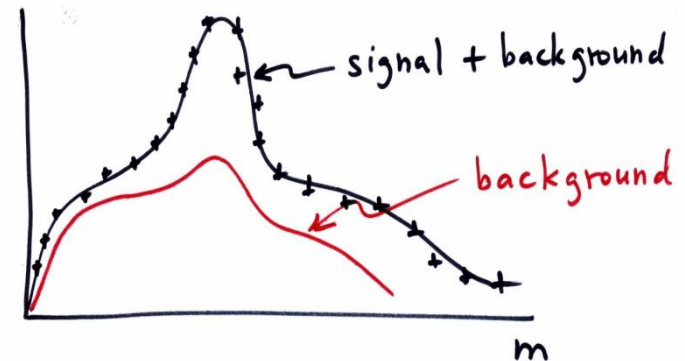
Difficulty is deciding how to introduce the additional parameters.

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.

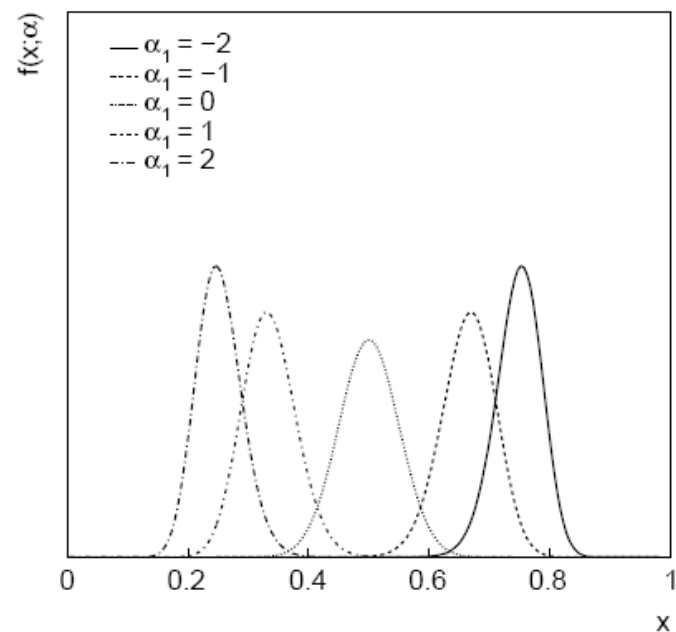
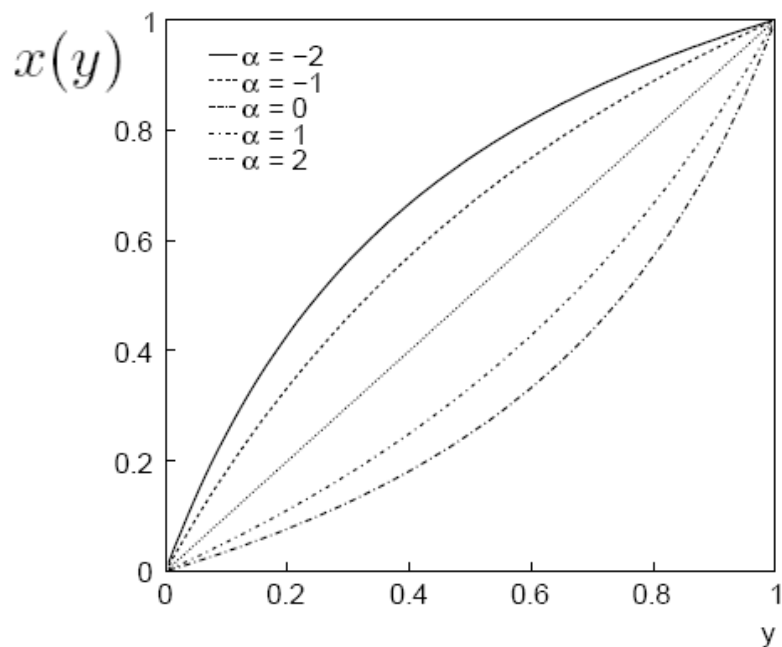
Both are broadened relative to original, reflecting the uncertainty connected with the nuisance parameter.

Step 2: uncertainty in shape

Key is to insert additional nuisance parameters into the model.

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

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

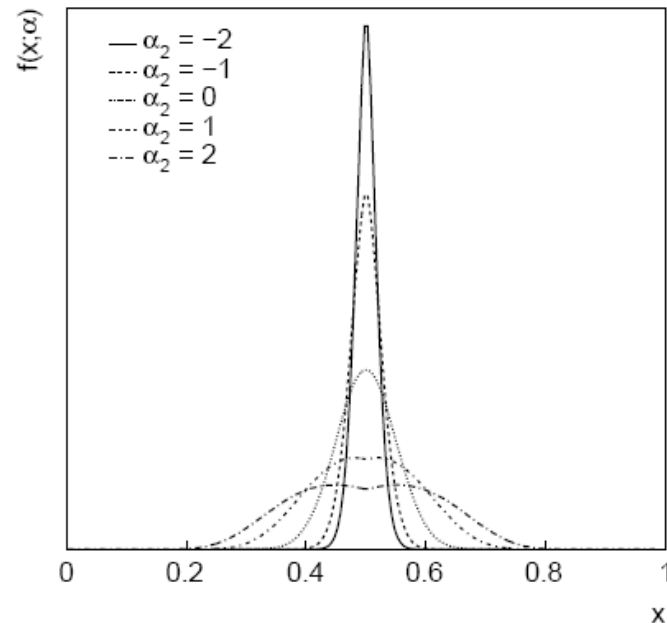
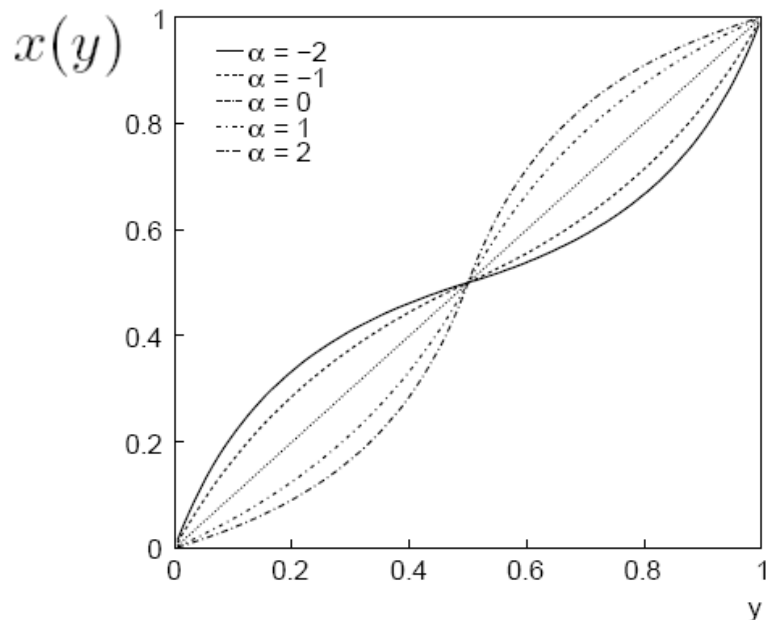


More uncertainty in shape

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:

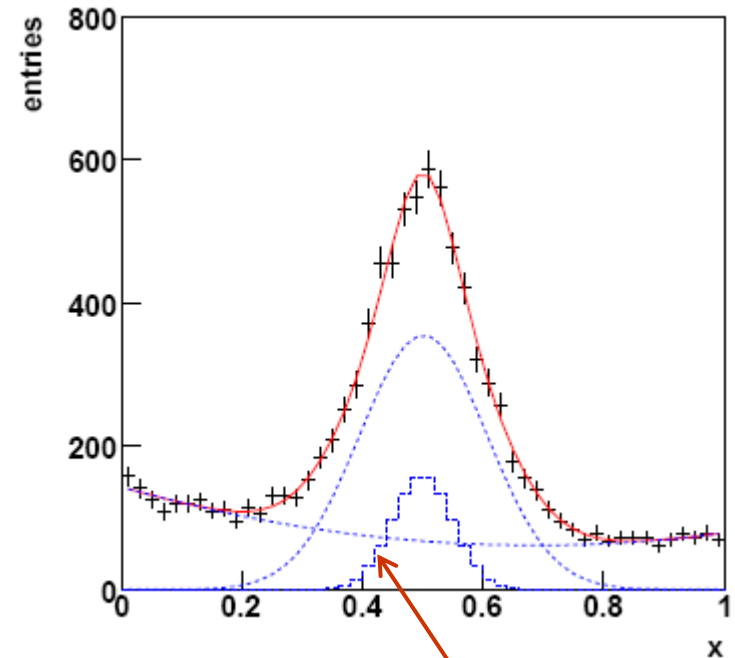
$$\mu_b = 0.5, \sigma_b = 0.05$$

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

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

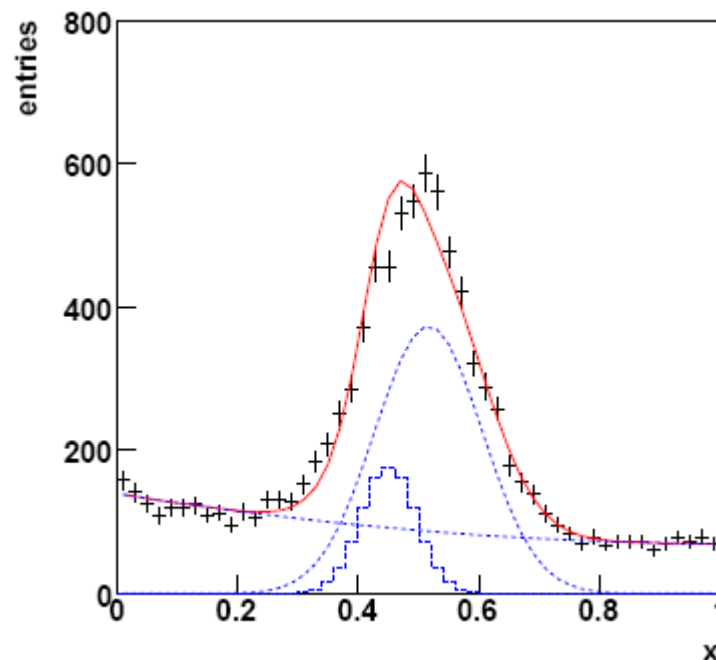
Now fitted values of signal parameters wrong, poor goodness-of-fit:

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

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

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

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)

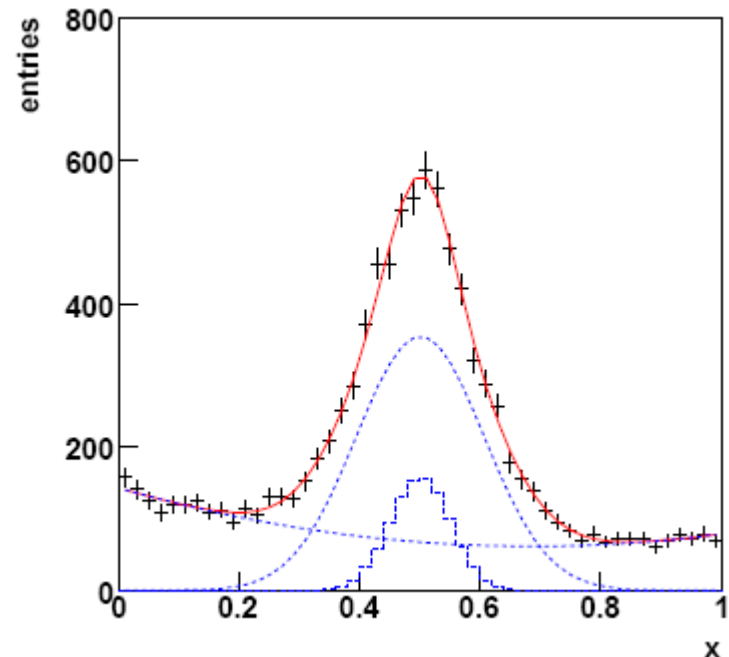
Result of fit is now “good”:

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

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

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

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).

Summary on systematics

Key to covering a systematic uncertainty is to include the appropriate nuisance parameters, constrained by all available info.

Enlarge model so that for at least one point in its parameter space, its difference from the truth is negligible.

In frequentist approach can use profile likelihood (similar with integrated product of likelihood and prior -- not discussed today).

Too many nuisance parameters spoils information about parameter(s) of interest.

In Bayesian approach, need to assign priors to (all) parameters.

Can provide important flexibility over frequentist methods.

Can be difficult to encode uncertainty in priors.

Exploit recent progress in Bayesian computation (MCMC).

Finally, when the LHC announces a 5 sigma effect, it's important to know precisely what the "sigma" means.

Extra slides

Summary on discovery

Current convention: p -value of background-only $< 2.9 \times 10^{-7}$ (5σ)

This should really depend also on other factors:

Plausibility of signal

Confidence in modeling of background

Can also use Bayes factor

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i) d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j) d\vec{\theta}_j}$$

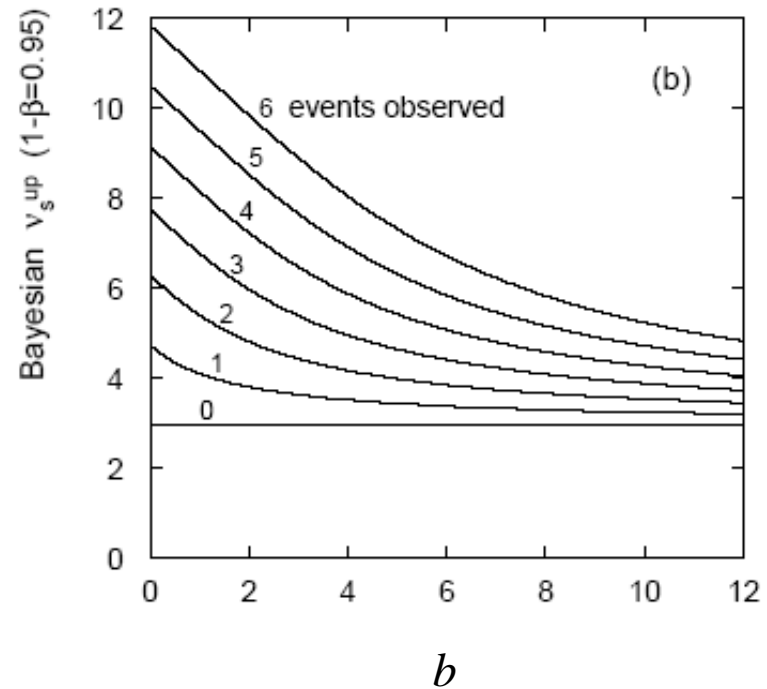
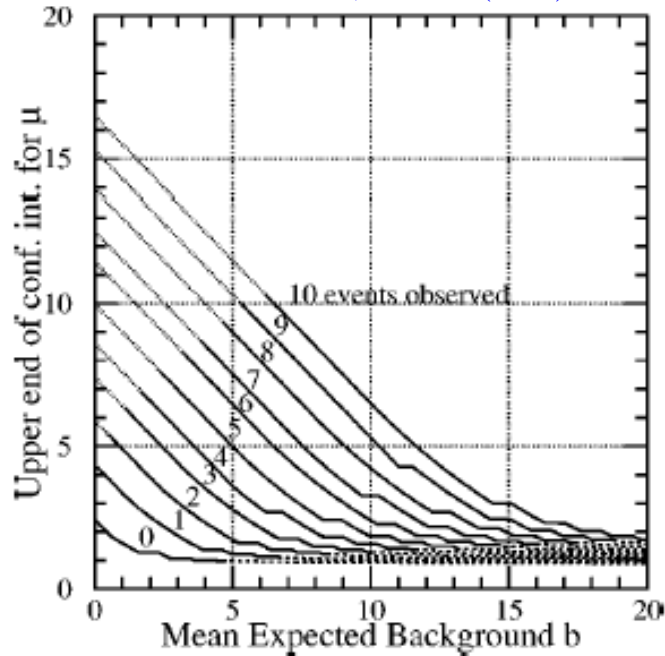
Should hopefully point to same conclusion as p -value.

If not, need to understand why!

As yet not widely used in HEP, numerical issues not easy.

Upper limit versus b

Feldman & Cousins, PRD 57 (1998) 3873



If $n = 0$ observed, should upper limit depend on b ?

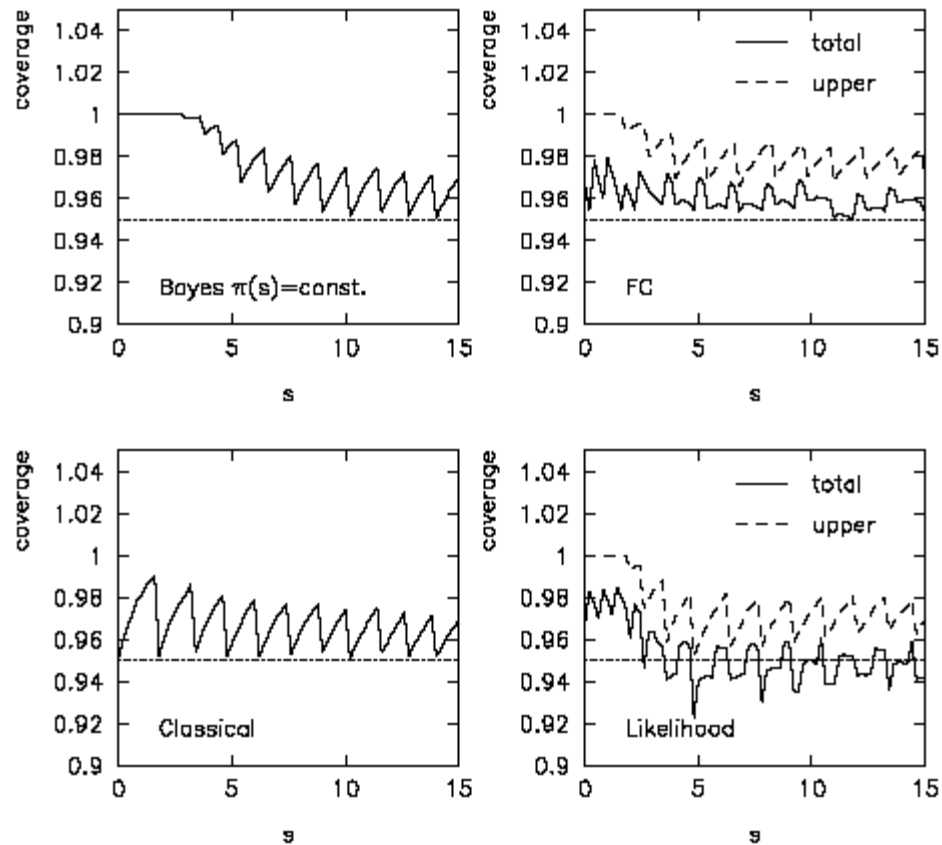
Classical: yes

Bayesian: no

FC: yes

Coverage probability of confidence intervals

Because of discreteness of Poisson data, probability for interval to include true value in general $>$ confidence level ('over-coverage')



Cousins-Highland method

Regard b as ‘random’, characterized by pdf $\pi(b)$.

Makes sense in Bayesian approach, but in frequentist model b is constant (although unknown).

A measurement b_{meas} is random but this is not the mean number of background events, rather, b is.

Compute anyway
$$P(n; s) = \int P(n; s, b) \pi_b(b) db$$

This would be the probability for n if Nature were to generate a new value of b upon repetition of the experiment with $\pi_b(b)$.

Now e.g. use this $P(n; s)$ in the classical recipe for upper limit at $\text{CL} = 1 - \beta$: $\beta = P(n \leq n_{\text{obs}}; s_{\text{up}})$

Result has hybrid Bayesian/frequentist character.

‘Integrated likelihoods’

Consider again signal s and background b , suppose we have uncertainty in b characterized by a prior pdf $\pi_b(b)$.

Define integrated likelihood as $L'(s) = \int L(s, b)\pi_b(b) db$, also called modified profile likelihood, in any case not a real likelihood.

Now use this to construct likelihood ratio test and invert to obtain confidence intervals.

Feldman-Cousins & Cousins-Highland (FHC²), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value s :

$$l(s) = \frac{L(n|s, b)}{L(n|\hat{s}, b)} \quad \text{where} \quad \hat{s} = \begin{cases} n - b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here \hat{s} is the ML estimator, note $0 \leq l(s) \leq 1$.

Define a **statistical test** for a hypothetical value of s :

Rejection region defined by low values of likelihood ratio.

Reject s if p -value = $P(l(s) \leq l_{\text{obs}} | s)$ is less than γ (e.g. $\gamma = 0.05$).

Confidence interval at $\text{CL} = 1 - \gamma$ is the set of s values not rejected.

Resulting intervals can be one- or two-sided (depending on n).

(Re)discovered for HEP by Feldman and Cousins,
Phys. Rev. D 57 (1998) 3873.

More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n \gg b$.

Usually one then quotes a measurement: $\hat{s} = n - b$, $\hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however, n isn't large enough to claim discovery, one sets a limit on s .

FC pointed out that if this decision is made based on n , then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on n .

But, suppose FC gives e.g. $0.1 < s < 5$ at 90% CL, p -value of $s=0$ still substantial. Part of upper-limit 'wasted'?

Sensitivity

Discovery:

Generate data under $s+b$ ($\mu = 1$) hypothesis;
Test hypothesis $\mu = 0 \rightarrow p\text{-value} \rightarrow Z$.

Exclusion:

Generate data under background-only ($\mu = 0$) hypothesis;
Test hypothesis $\mu = 1$.
If $\mu = 1$ has $p\text{-value} < 0.05$ exclude m_H at 95% CL.

Presence of nuisance parameters leads to broadening of the profile likelihood, reflecting the loss of information, and gives appropriately reduced discovery significance, weaker limits.

Sensitivity

Discovery:

Generate data under $s+b$ ($\mu = 1$) hypothesis;

Test hypothesis $\mu = 0 \rightarrow p\text{-value} \rightarrow Z$.

Exclusion:

Generate data under background-only ($\mu = 0$) hypothesis;

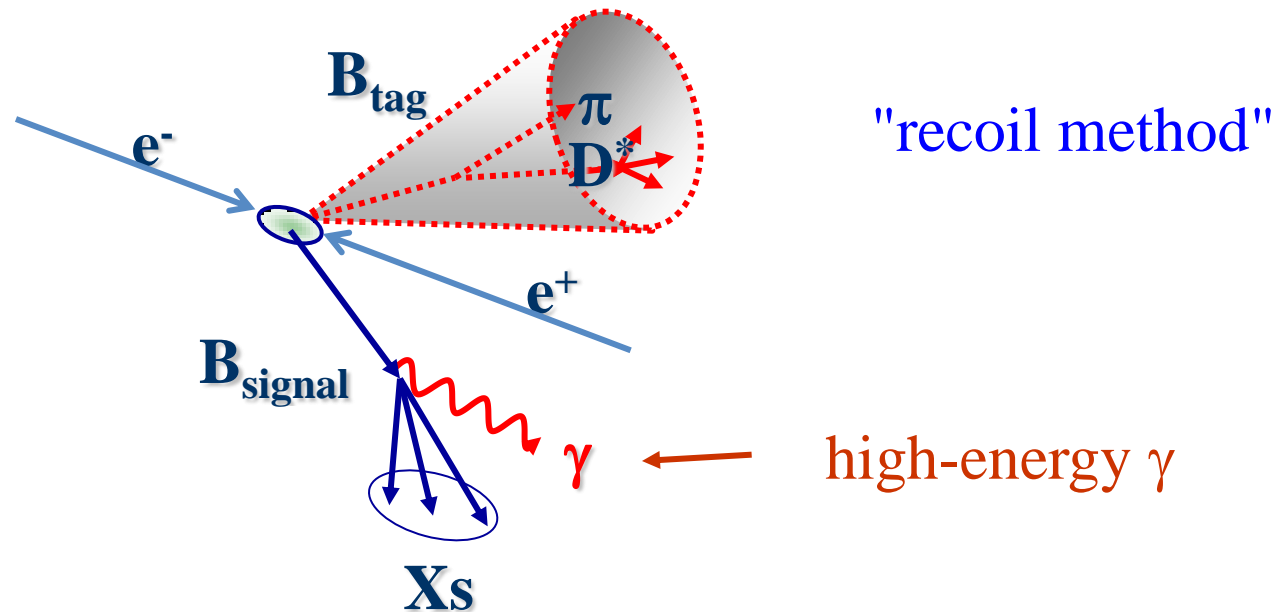
Test hypothesis $\mu = 1$.

If $\mu = 1$ has $p\text{-value} < 0.05$ exclude m_H at 95% CL.

Presence of nuisance parameters leads to broadening of the profile likelihood, reflecting the loss of information, and gives appropriately reduced discovery significance, weaker limits.

Fit example: $b \rightarrow s\gamma$ (BaBar)

B. Aubert et al. (BaBar), Phys. Rev. D 77, 051103(R) (2008).



Decay of one B fully reconstructed (B_{tag}).

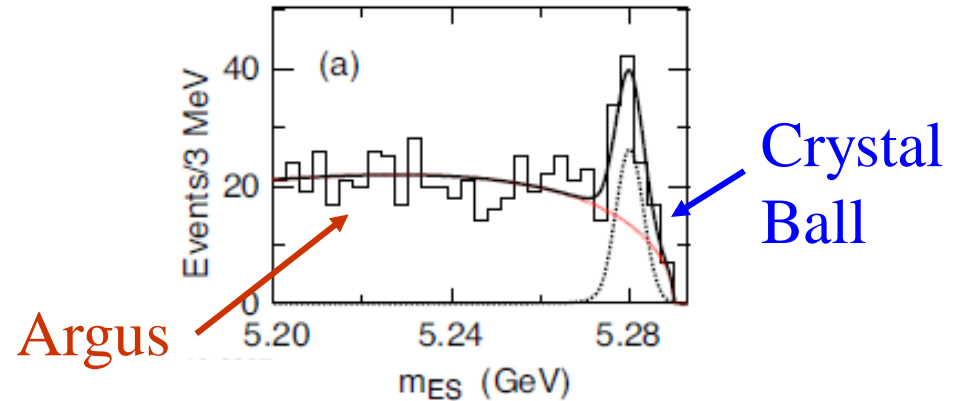
Look for high-energy γ from rest of event.

Signal and background yields from fit to m_{ES} in bins of E_γ .

$$m_{ES} = \sqrt{E_{\text{beam}}^{*2} - p_{\text{tag}}^2} \quad (\approx m_B \text{ for signal})$$

Fitting m_{ES} distribution for $b \rightarrow s\gamma$

Fit m_{ES} distribution using superposition of Crystal Ball and Argus functions:



$$c(m; \alpha, \beta, \mu, \sigma) = \begin{cases} N e^{-(m-\mu)^2/2\sigma^2} & (m - \mu)/\sigma > -\alpha, \\ N \left(\frac{\beta}{|\alpha|} - |\alpha| - \frac{m-\mu}{\sigma} \right)^{-\beta} \left(\frac{\beta}{|\alpha|} \right)^\beta e^{-\alpha^2/2} & \text{otherwise.} \end{cases}$$

$$a(m; \xi) = \begin{cases} N m \sqrt{1 - \left(\frac{m}{m_{\max}} \right)^2} \exp \left[-\xi \left(1 - \left(\frac{m}{m_{\max}} \right)^2 \right) \right] & 0 < m \leq m_{\max}, \\ 0 & \text{otherwise,} \end{cases}$$

log-likelihood: $\ln L(\nu_c, \nu_a, \alpha, \beta, \mu, \sigma, \xi) = \sum_{i=1}^N (n_i \ln \nu_i - \nu_i)$

↑
↑
↑
↑

rates
shapes
obs./pred. events in i th bin

Simultaneous fit of all m_{ES} distributions

Need fits of m_{ES} distributions in 14 bins of E_γ :

At high E_γ , not enough events to constrain shape, so combine all E_γ bins into global fit:

$$\ln L(\vec{\nu}_c, \vec{\nu}_a, \vec{\alpha}, \vec{\beta}, \vec{\mu}, \vec{\sigma}, \vec{\xi}) = \sum_{i=1}^M \ln L(\nu_{c,i}, \nu_{a,i}, \alpha_i, \beta_i, \mu_i, \sigma_i, \xi_i)$$

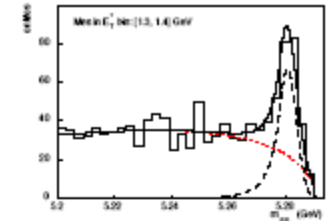
Shape parameters could vary (smoothly) with E_γ .

So make Ansatz for shape parameters such as

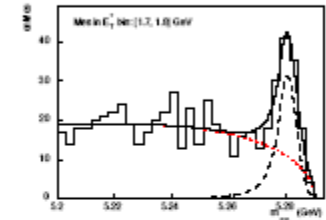
$$\alpha(E) = \alpha_0 + \alpha_1 E + \alpha_2 E^2 + \dots$$

Start with no energy dependence, and include one by one more parameters until data well described.

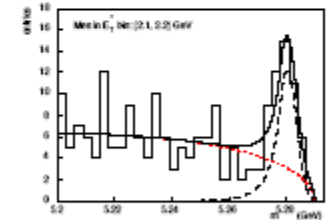
1.3 GeV < E_γ < 1.4 GeV



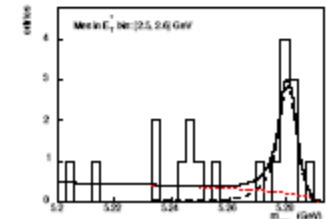
1.7 GeV < E_γ < 1.8 GeV



2.1 GeV < E_γ < 2.2 GeV



2.5 GeV < E_γ < 2.6 GeV



Finding appropriate model flexibility

Here for Argus ξ parameter, linear dependence gives significant improvement; fitted coefficient of linear term -10.7 ± 4.2 .

fit option	χ^2	degrees of freedom
(1) no E dependence	389.70	387
(2) linear for Argus ξ	386.22	386
(3) quadratic for Argus ξ	385.61	385
(4) linear for ξ and α	386.29	385
(5) linear for ξ and σ	386.42	385
(6) linear for ξ and μ	386.12	385
(7) linear for ξ, α, σ, μ	385.59	383

← $\chi^2(1) - \chi^2(2) = 3.48$
 p -value of (1) = 0.062
→ data want extra par.

D. Hopkins, PhD thesis, RHUL (2007).

Inclusion of additional free parameters (e.g., quadratic E dependence for parameter ξ) do not bring significant improvement.

So including the additional energy dependence for the shape parameters converts the systematic uncertainty into a statistical uncertainty on the parameters of interest.