Systematic uncertainties in statistical data analysis for particle physics

DESY Seminar

Hamburg, 31 March, 2009

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

Outline

Preliminaries

Role of probability in data analysis (Frequentist, Bayesian)

Systematic errors and nuisance parameters

A simple fitting problem

Frequentist solution / Bayesian solution

When does $\sigma_{tot}^2 = \sigma_{stat}^2 + \sigma_{sys}^2$ make sense?

Systematic uncertainties in a search

Example of search for Higgs (ATLAS)

Examples of nuisance parameters in fits

$b \to s\gamma$ with recoil method (BaBar)

Towards a general strategy for nuisance parameters

Conclusions

Data analysis in HEP

Particle physics experiments are expensive

e.g. LHC, ~ $\$10^{10}$ (accelerator and experiments)

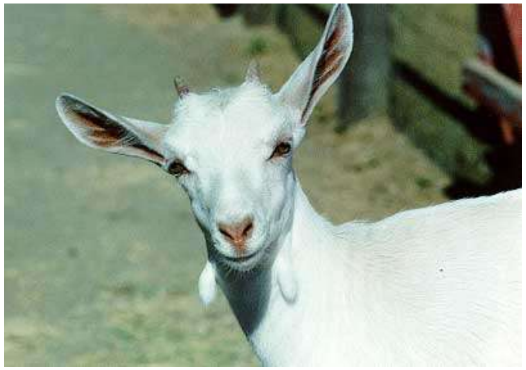
the competition is intense

(ATLAS vs. CMS) vs. Tevatron

and the stakes are high:

4 sigma effect

5 sigma effect



So there is a strong motivation to know precisely whether one's signal is a 4 sigma or 5 sigma effect.

Frequentist vs. Bayesian approaches

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

The preferred hypotheses (theories, models, parameter values, ...) are those for which our observations would be considered 'usual'.

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability).

Use Bayes' theorem to relate (posterior) probability for hypothesis H given data $\boldsymbol{x}$ to probability of $\boldsymbol{x}$ given H (the likelihood):

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

Need prior probability, $\pi(H)$, i.e., before seeing the data.

Statistical vs. systematic errors

Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

Systematic errors:

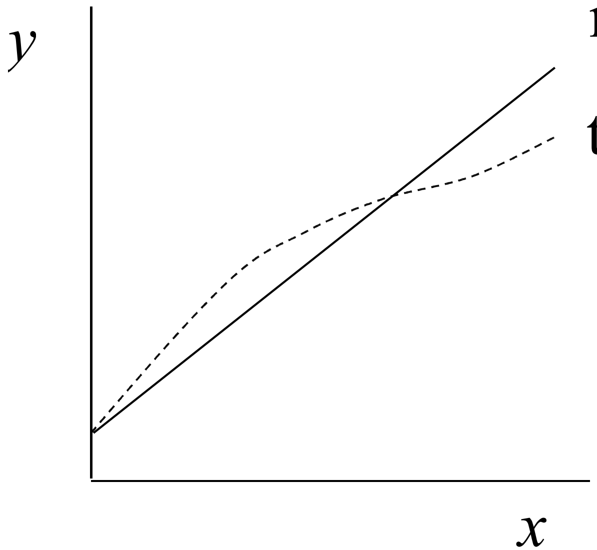
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling of measurement apparatus.

Usually taken to mean the sources of error do not vary upon repetition of the measurement. Often result from uncertain value of calibration constants, efficiencies, etc.

Systematic errors and nuisance parameters

Model prediction (including e.g. detector effects) never same as "true prediction" of the theory:



model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

Model can be made to approximate better the truth by including more free parameters.

systematic uncertainty ↔ nuisance parameters

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \ldots, n$.

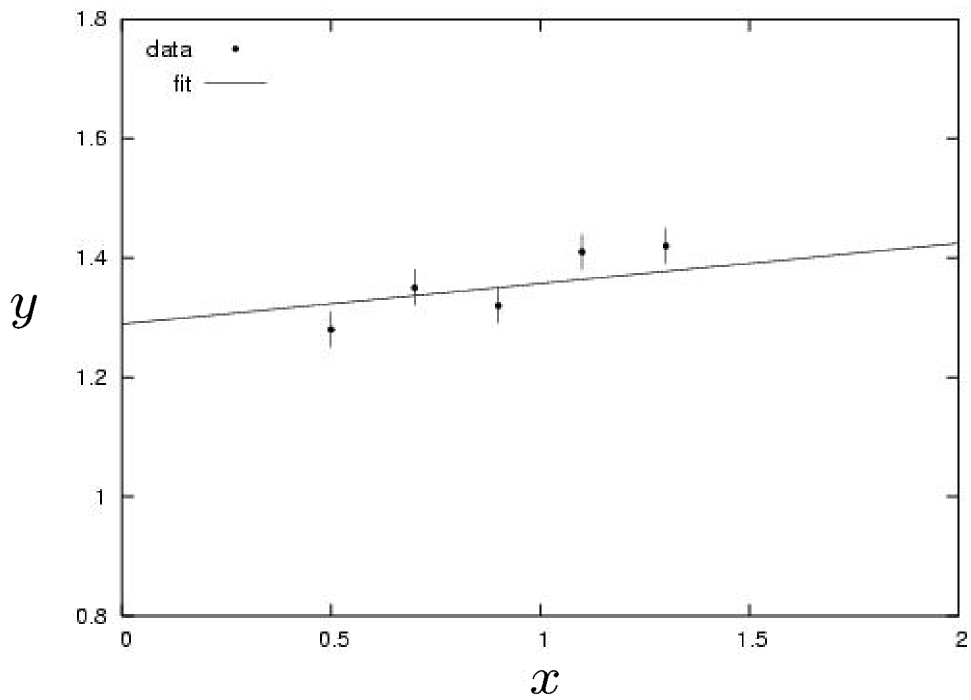
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Frequentist approach

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

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

Standard deviations from

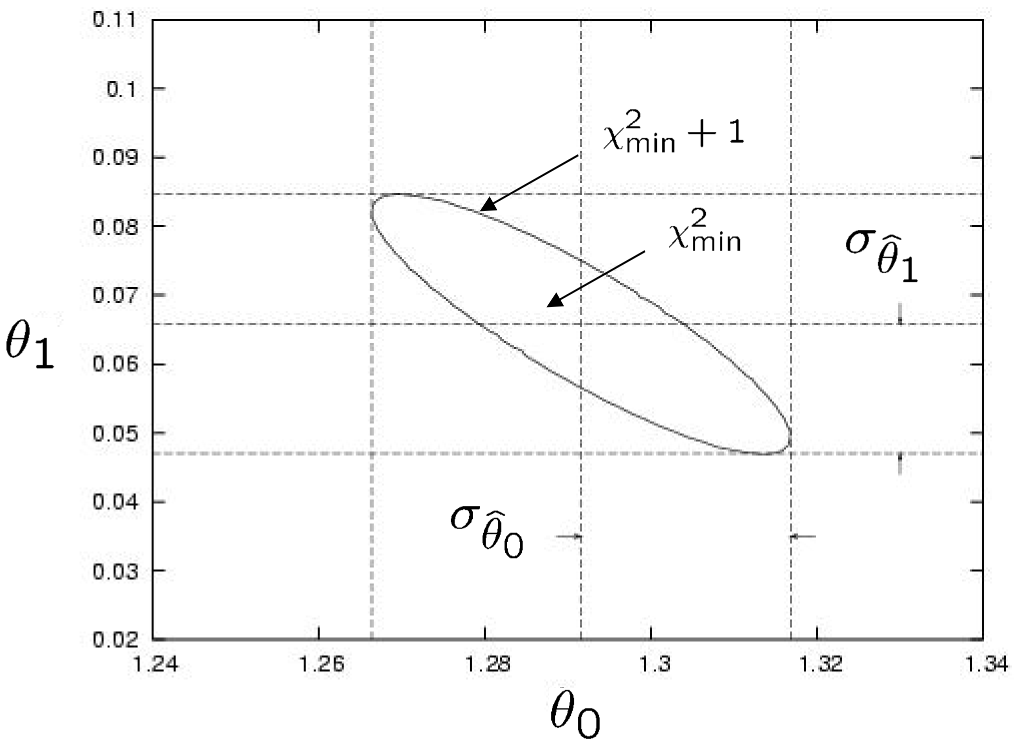
tangent lines to contour

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

Correlation between

$\hat{\theta}_0, \ \hat{\theta}_1$ causes errors

to increase.

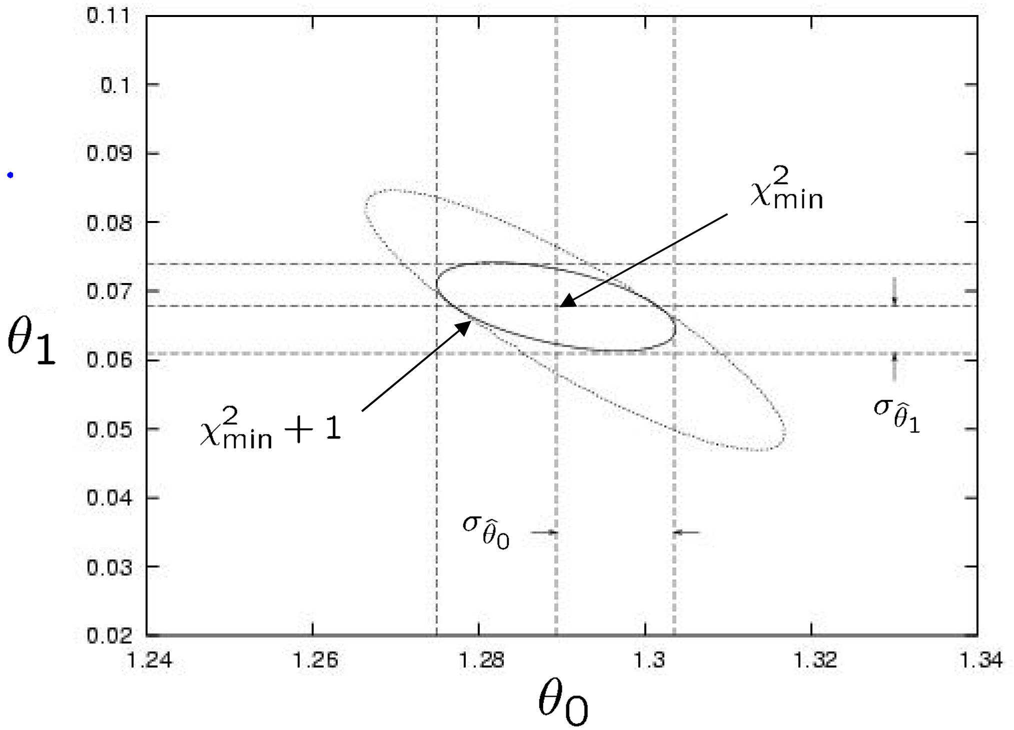


Frequentist case with a measurement t_1 of θ_1

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

The information on θ_1

improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\, \pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

← reflects 'prior ignorance', in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

← based on previous measurement

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) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

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

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

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

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

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

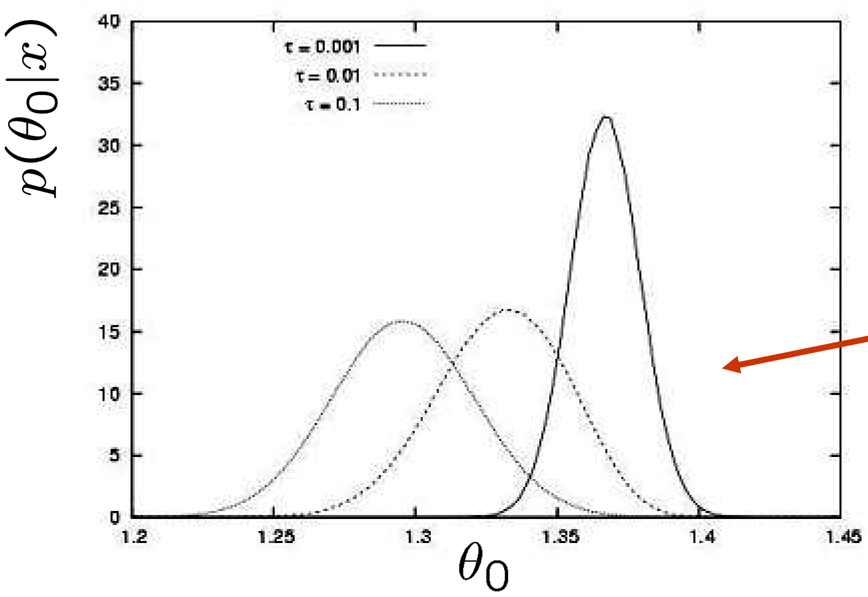
Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_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 θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

A more general fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: $V_{ij}^{\text{stat}}, V_{ij}^{\text{sys}}$.

Predicted value: $\mu(x_i;\theta)$, expectation value $E[y_i] = \mu(x_i;\theta) + b_i$

control variable parameters bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing $L(\theta)$ » $e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $$L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T V_{\mathrm{stat}}^{-1}(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$$

$$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\vec{b}^T V_{\mathrm{sys}}^{-1}\vec{b}\right]$$

$$\pi_\theta(\theta) \sim \mathrm{const.}$$

Joint probability for all parameters

and use Bayes' theorem: $$p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$$

To get desired probability for θ, integrate (marginalize) over $\boldsymbol{b}$:

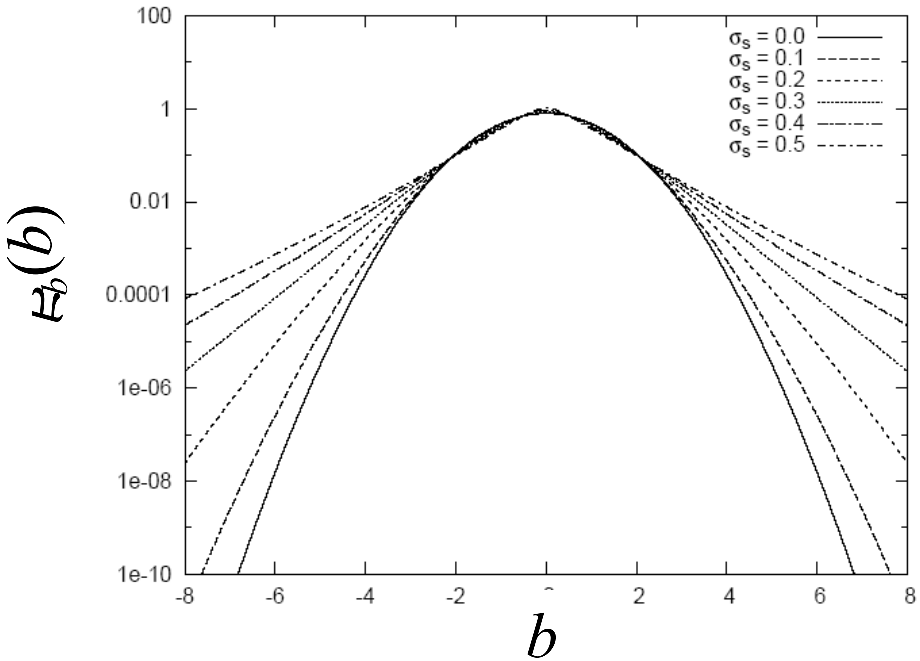
$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\mathrm{min}} + 1$. (Back where we started!)

Alternative priors for systematic errors

Gaussian prior for the bias b often not realistic, especially if one considers the "error on the error". Incorporating this can give a prior with longer tails:

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\rm sys}} \exp\left[-\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\rm sys})^2} \right] \pi_s(s_i)\, ds_i$$



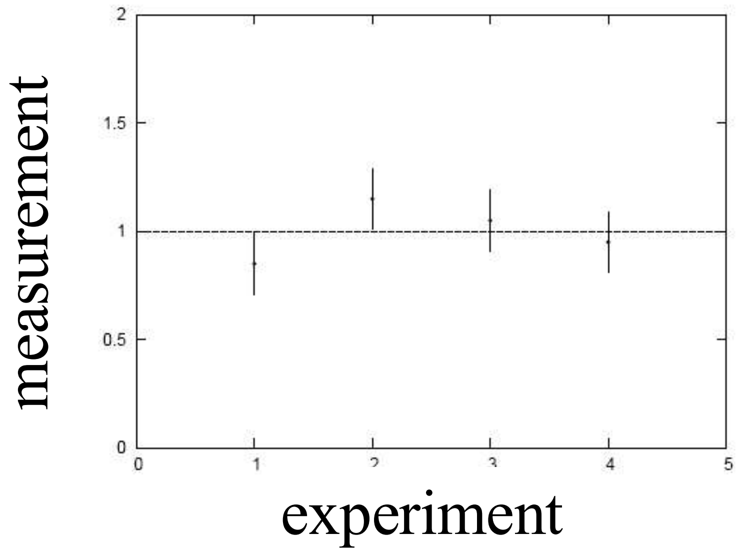
Represents 'error on the error'; standard deviation of $\pi_s(s)$ is σ_s.

A simple test

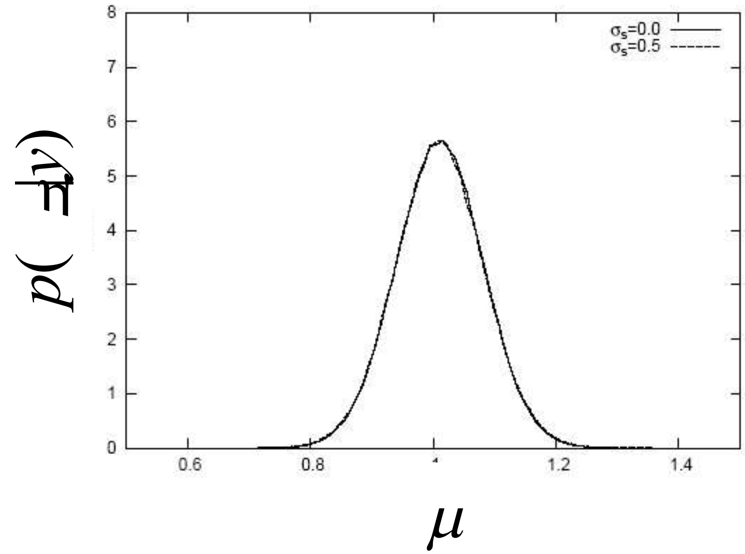
Suppose fit effectively averages four measurements.

Take $\sigma_{sys} = \sigma_{stat} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



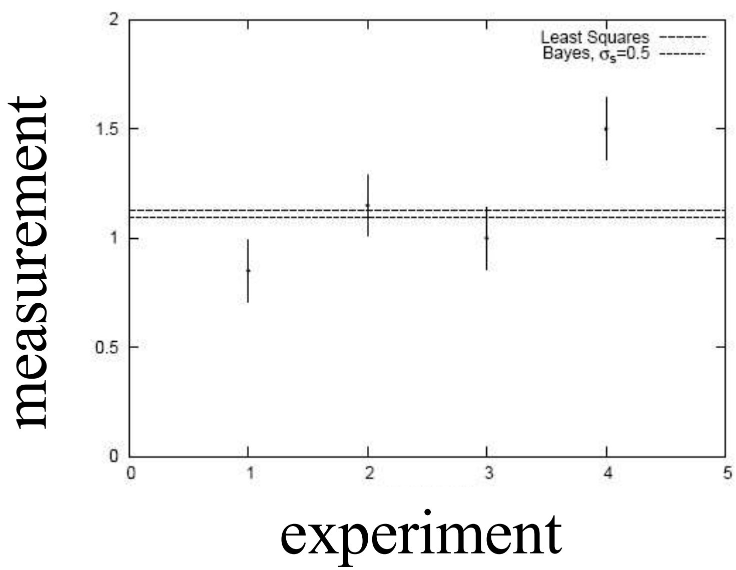
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$$\sigma_s = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$$

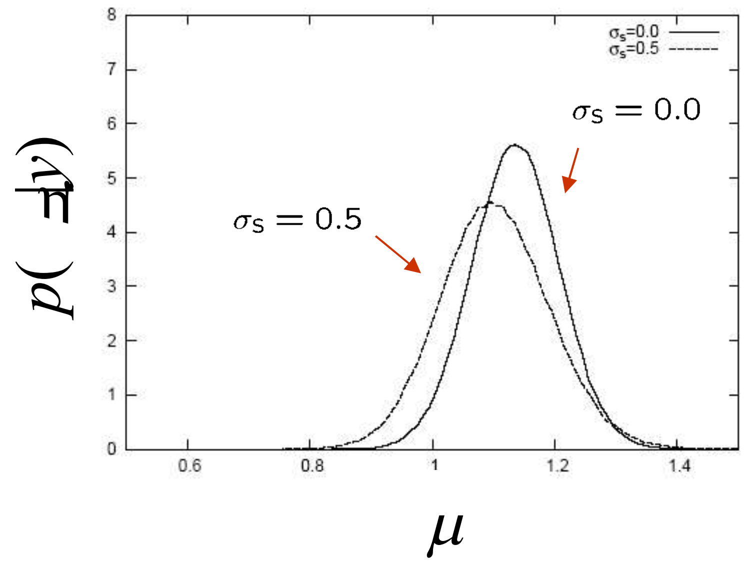
$$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$$

→ Bayesian fit less sensitive to outlier.

(See also D'Agostini 1999; Dose & von der Linden 1999)

Example of systematics in a search

Combination of Higgs 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):

$H \to \gamma\gamma$

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

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

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

Used profile likelihood method for systematic uncertainties:

background rates, signal & background shapes.

Statistical model for Higgs search

Bin i of a given channel has n_i events, expectation value is

$$E[n_i] = \mu L \varepsilon_i \sigma_i \mathcal{B} + b_i \equiv \mu s_i + b_i$$

μ is global strength parameter, common to all channels. $\mu = 0$ means background only, $\mu = 1$ is SM hypothesis.

Expected signal and background are:

$$s_i = s_{\rm tot} \int_{{\rm bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx \,,$$

$$b_i = b_{\rm tot} \int_{{\rm bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx$$

$b_{\rm tot}$, $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_b$ are nuisance parameters

The likelihood function

The single-channel likelihood function uses Poisson model for events in signal and control histograms:

data in signal histogram

data in control histogram

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

here signal rate is only parameter of interest

$\boldsymbol{\theta}$ represents all nuisance parameters, e.g., background rate, shapes

There is a likelihood $L_i(\mu,\boldsymbol{\theta}_i)$ for each channel, $i = 1, \ldots, N$.

The full likelihood function is $L(\mu, \boldsymbol{\theta}) = \prod_i L_i(\mu, \boldsymbol{\theta}_i)$

Profile likelihood ratio

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

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

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

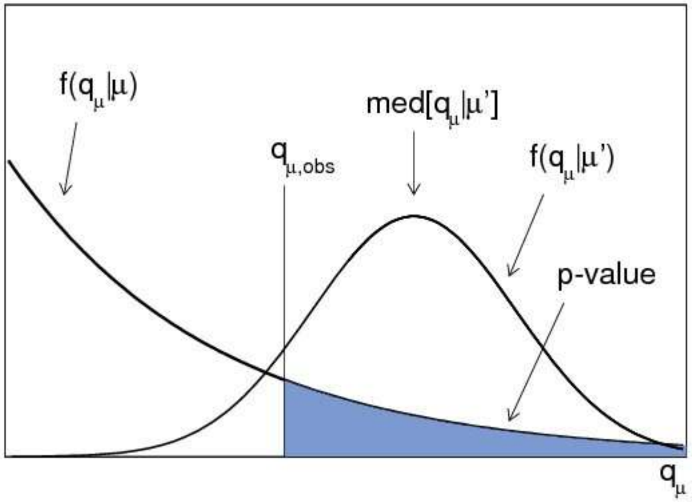
Distribution of q_μ under assumption of μ related to chi-square

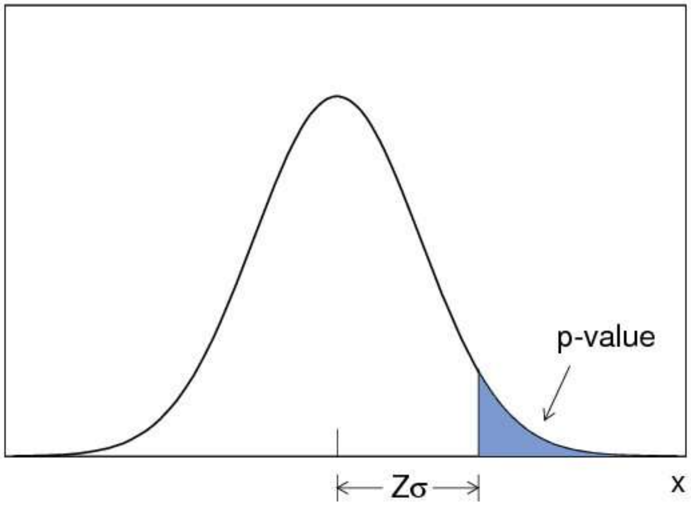
(Wilks' theorem, here approximation valid for roughly $L > 2\ \mathrm{fb}^{-1}$):

$$f(q_\mu|\mu) \approx \frac{1}{2} f_{\chi^2_1}(q_\mu) + \frac{1}{2}\delta(q_\mu)$$

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:

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

Sensitivity

Discovery:

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

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

Exclusion:

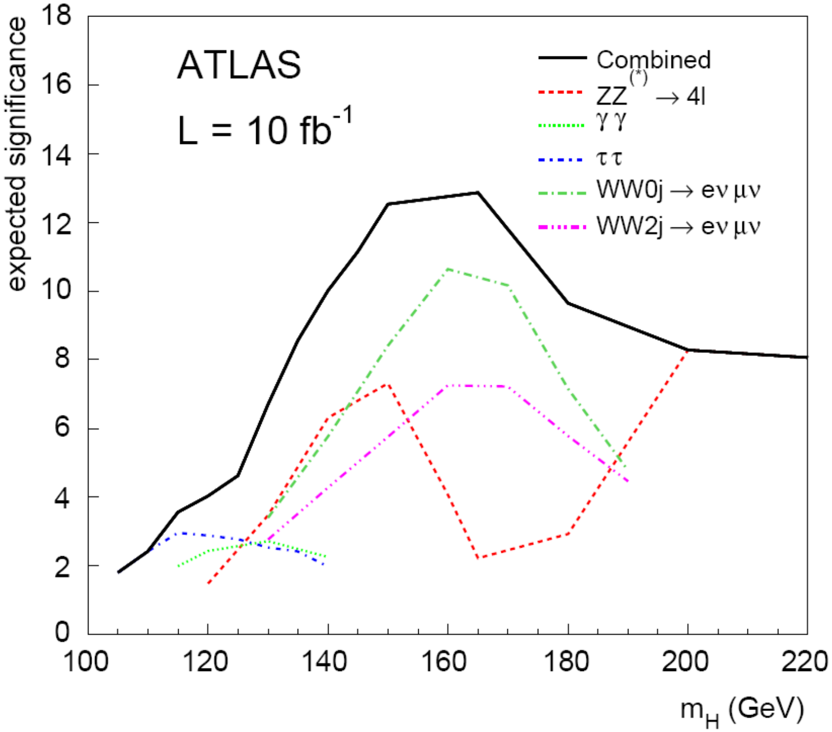
Generate data under background-only ($\mu = 0$) hypothesis;

Test hypothesis $\mu = 1$.

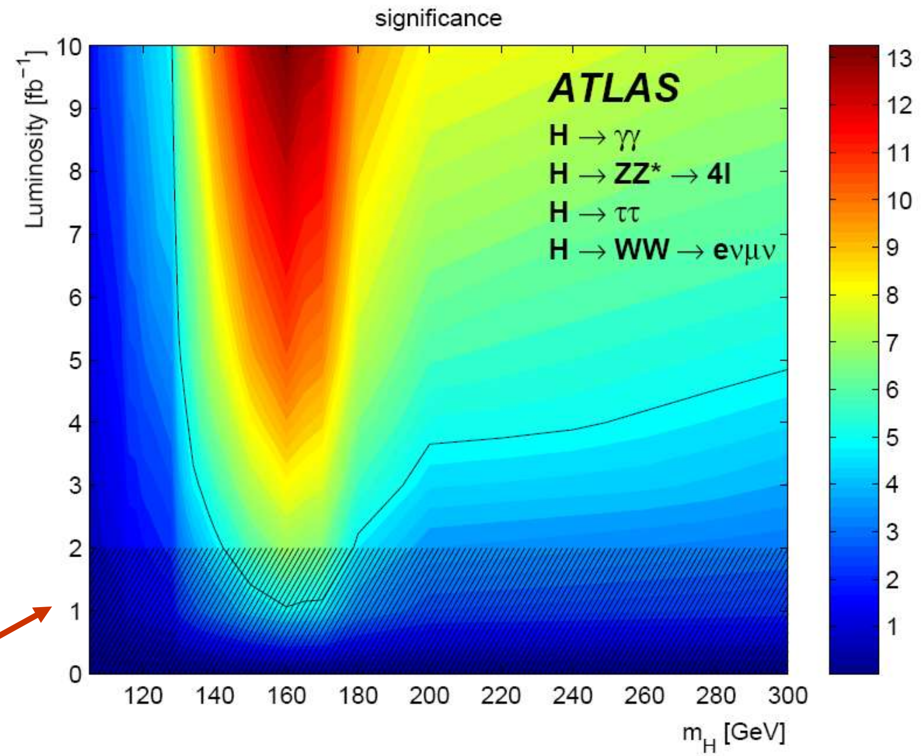
If $\mu = 1$ has p-value < 0.05 exclude m_{H} at 95% CL.

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

Combined discovery significance



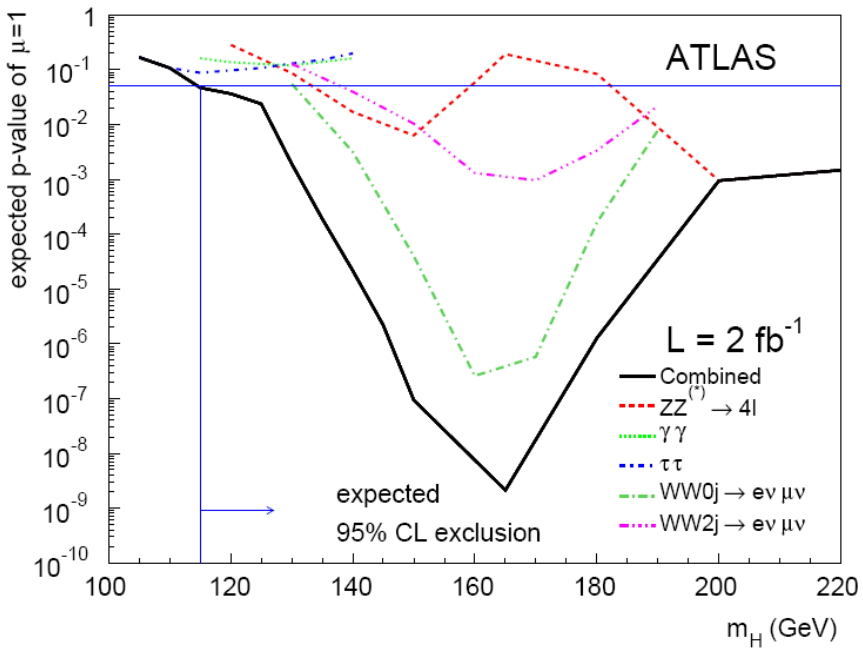
Discovery signficance (in colour) vs. L, m_H:

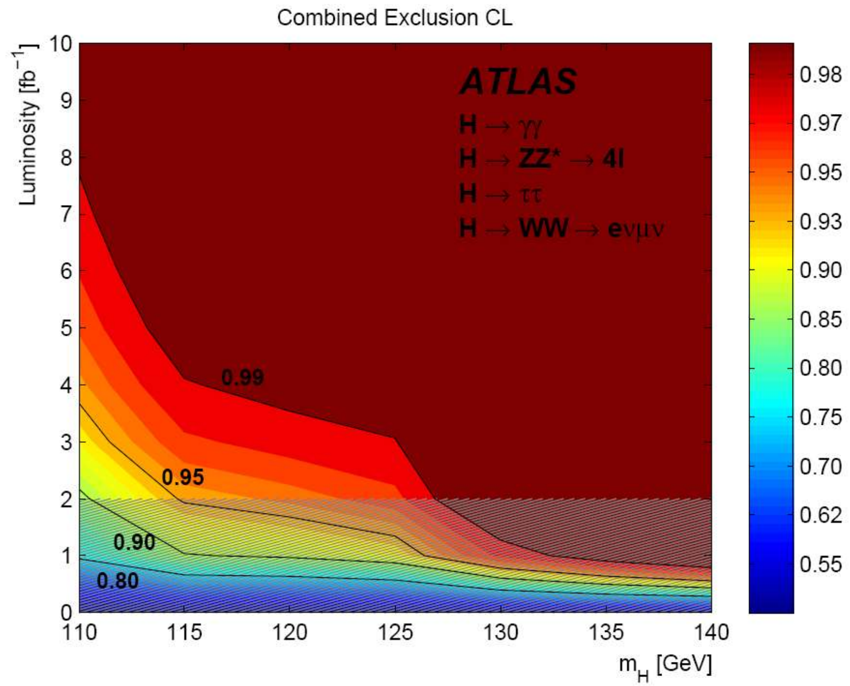


Approximations used here not always accurate for $L < 2\ \mathrm{fb}^{-1}$ but in most cases conservative.

Combined 95% CL exclusion limits

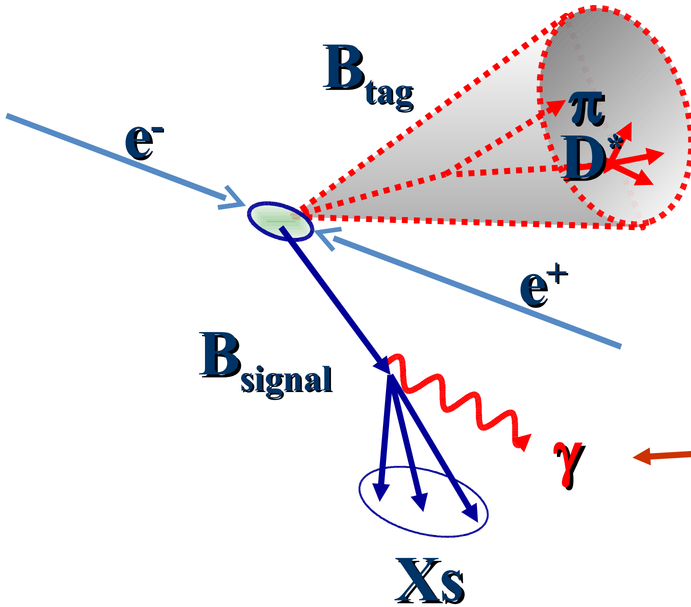
1 – p-value of m_H
(in colour) vs. L, m_H:





Fit example: b → sγ (BaBar)

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



"recoil method"

high-energy γ

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

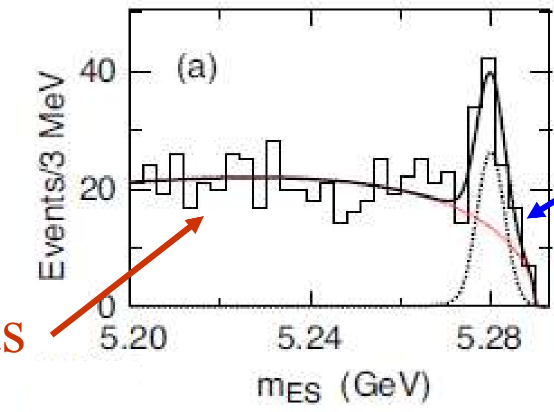
Look for high-energy γ from rest of event.

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

$$m_{\mathrm{ES}} = \sqrt{E^{*2}_{\mathrm{beam}} - p^2_{\mathrm{tag}}} \quad (\approx m_{\mathrm{B}} \text{ for signal})$$

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

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



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

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

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 (ν_c, ν_a) shapes ($\alpha, \beta, \mu, \sigma, \xi$) obs./pred. events in ith bin (n_i, ν_i)

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:

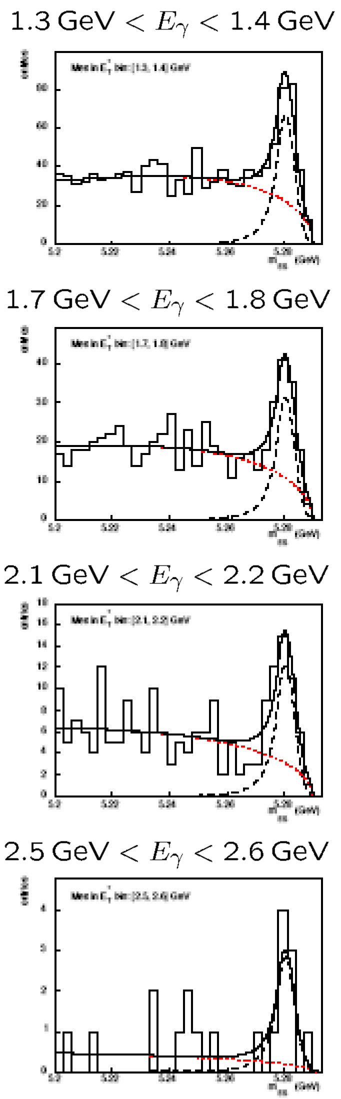
$$\ln L(\vec{\nu}_{\mathrm{c}}, \vec{\nu}_{\mathrm{a}}, \vec{\alpha}, \vec{\beta}, \vec{\mu}, \vec{\sigma}, \vec{\xi}) = \sum_{i=1}^{M} \ln L(\nu_{\mathrm{c},i}, \nu_{\mathrm{a},i}, \alpha_i, \beta_i, \mu_i, \sigma_i, \xi_i)$$

Shape parameters could vary (smoothly) with E_γ.

So make Ansatz for shape parameters such as

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

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



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

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

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

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.

Towards a general strategy (frequentist)

In progress together with: S. Caron, S. Horner, J. Sundermann, E. Gross, O Vitells, A. Alam

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.

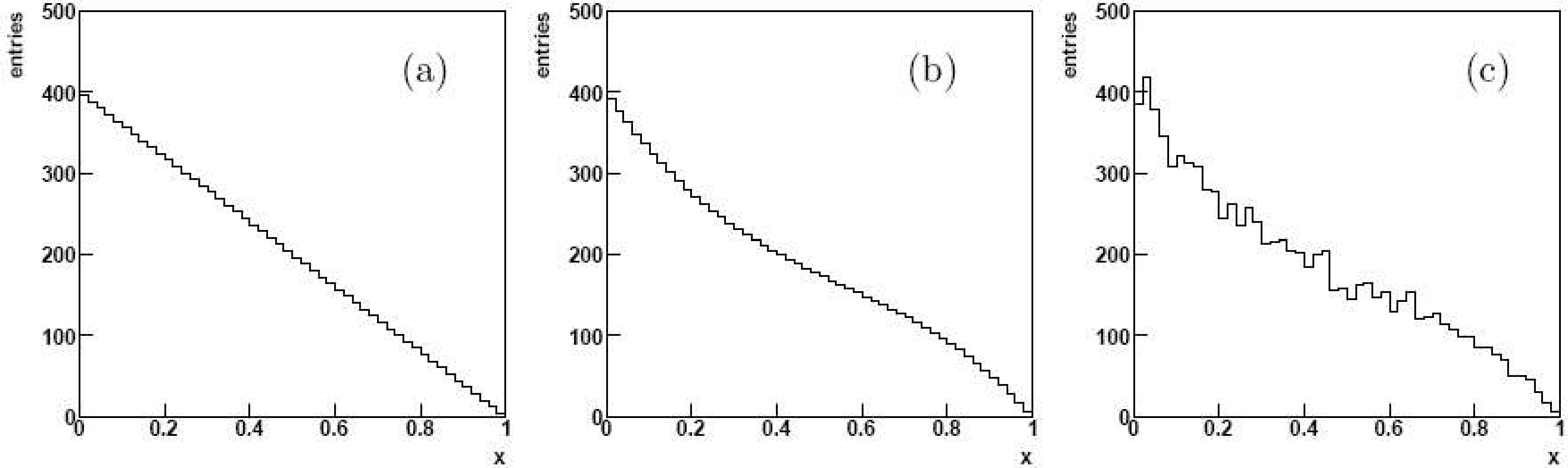
Difficulty is deciding how to introduce the additional parameters.

A simple example

0th order model

True model (Nature)

Data



The naive model (a) could have been e.g. from MC (here statistical errors suppressed; point is to illustrate how to incorporate systematics.)

Comparing model vs. data

Model number of entries n_i in ith bin as ~Poisson(ν_i)

$$P(\mathbf{n};\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

In the example shown, the model and data clearly don't agree well.

To compare, use e.g.

$$\chi^2_{\mathrm{P}} = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

Will follow chi-square distribution for N dof for sufficiently large n_i.

Model-data comparison with likelihood ratio

This is very similar to a comparison based on the likelihood ratio

$$\lambda(\boldsymbol{\nu}) = \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})}$$

where $L(\boldsymbol{\nu}) = P(\boldsymbol{n};\boldsymbol{\nu})$ is the likelihood and the hat indicates the ML estimator (value that maximizes the likelihood).

Here easy to show that $\hat{\nu}_i = n_i$

Equivalently use logarithmic variable

$$q_{\boldsymbol{\nu}} = -2 \ln \lambda(\boldsymbol{\nu}) = 2 \sum_{i=1}^{N} \left(n_i \ln \frac{n_i}{\nu_i} + \nu_i - n_i \right)$$

If model correct, $q_\nu \sim$ chi-square for N degrees of freedom.

p-values

Using either χ^2_P or $q_{\boldsymbol{\nu}}$, state level of data-model agreement by giving the p-value: the probability, under assumption of the model, of obtaining an equal or greater incompatibility with the data relative to that found with the actual data:

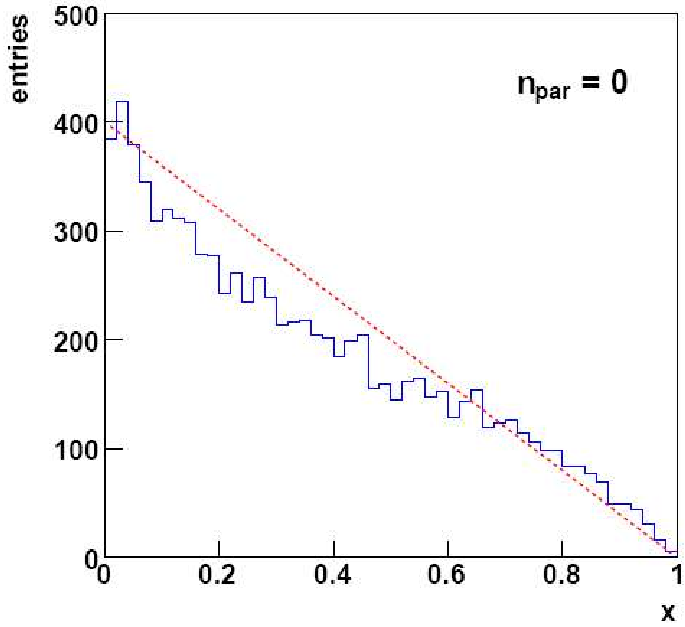
$$p = \int_{q_{\boldsymbol{\nu},\text{obs}}}^{\infty} f_{\chi^2}(z; N)\, dz$$

where (in both cases) the integrand is the chi-square distribution for N degrees of freedom,

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

Comparison with the 0th order model

The 0th order model gives $q_\nu = 258.8,\ p = 6 \times 10^{-30}$



Enlarging the model

Here try to enlarge the model by multiplying the 0th order distribution by a function *s*:

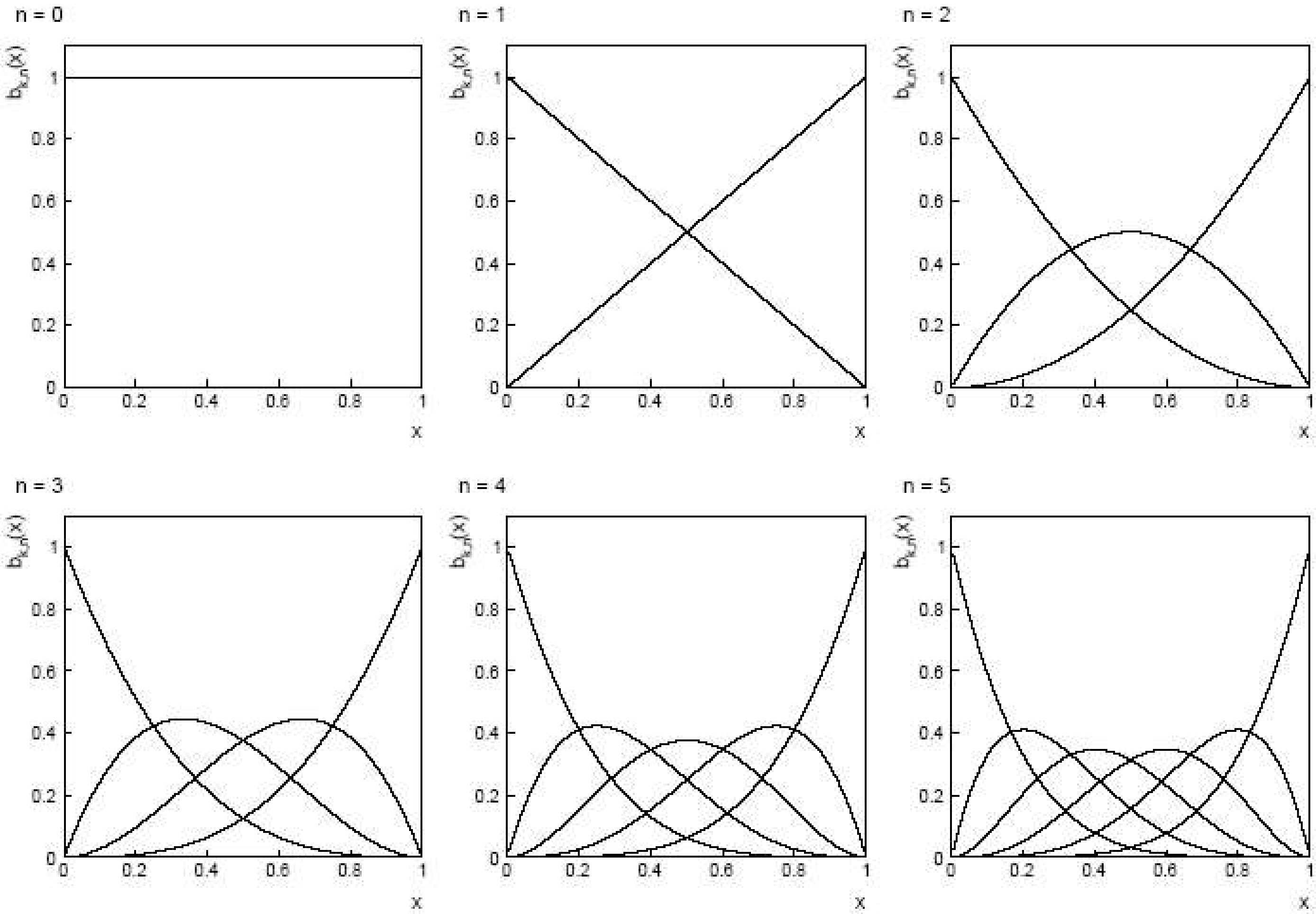
$$\nu_i \to \nu_i s(x_i; \boldsymbol{\theta})$$

where $s(x)$ is a linear superposition of Bernstein basis polynomials of order m:

$$s(x) = \sum_{k=0}^{m} \beta_k b_{k,m}(x)$$

$$b_{k,m}(x) = \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k}$$

Bernstein basis polynomials



Using increasingly high order for the basis polynomials gives an increasingly flexible function.

Enlarging the parameter space

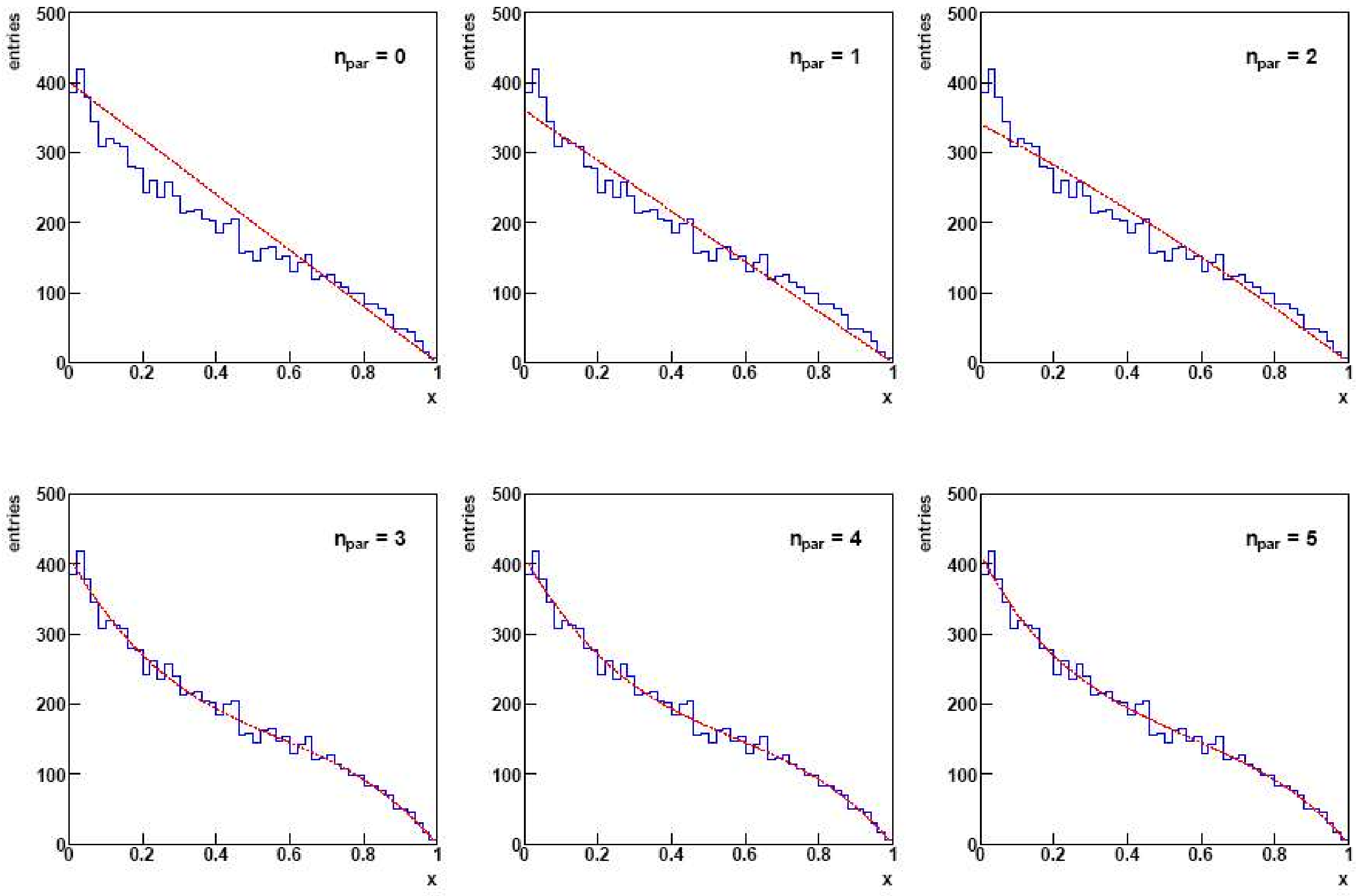
Using increasingly high order for the basis polynomials gives an increasingly flexible function.

At each stage compare the *p*-value to some threshold, e.g., 0.1 or 0.2, to decide whether to include the additional parameter.

Now iterate this procedure, and stop when the data do not require addition of further parameters based on the likelihood ratio test. (And overall goodness-of-fit should also be good.)

Once the enlarged model has been found, simply include it in any further statistical procedures, and the statistical errors from the additional parameters will account for the systematic uncertainty in the original model.

Fits using increasing numbers of parameters



Stop here

Deciding appropriate level of flexibility

When p-value exceeds ~0.1 to 0.2, fit is good enough.

n_{par}	q_ν	p_ν	q	p
0	258.8	6.1×10^{-30}	98.9	2.6×10^{-23}
1	159.9	1.1×10^{-13}	15.4	8.9×10^{-05}
2	144.5	1.3×10^{-11}	112.0	3.5×10^{-26}
3	32.5	0.95	0.0013	0.97
4	32.5	0.93	0.26	0.61
5	32.2	0.92	0.37	0.54

← Stop here (row n_{par} = 3)

↑ (p_ν) says whether data prefer additional parameter

↑ (p) says whether data well described overall

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 and conclusions

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

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.

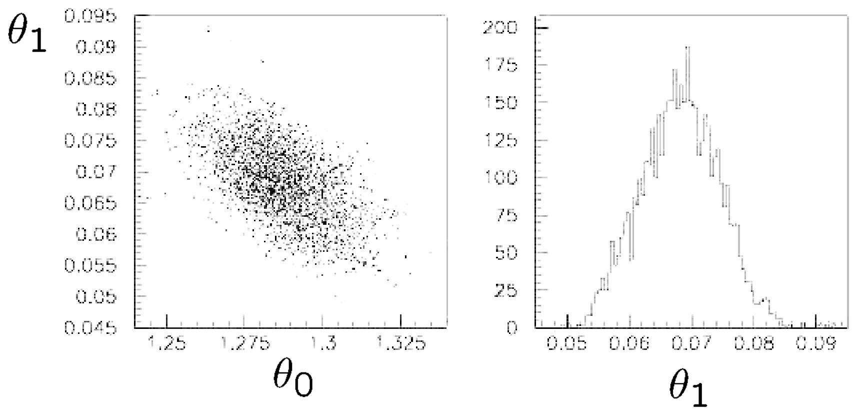
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

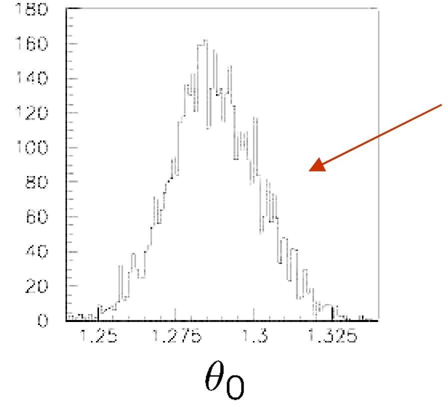
cannot use for many applications, e.g., detector MC;
effective stat. error greater than naive $\sqrt{n}$.

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

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?)

MCMC basics: Metropolis-Hastings algorithm

Goal: given an *n*-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

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 $u \sim \mathrm{Uniform}[0, 1]$

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 naive $\sqrt{n}$.

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.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a “burn-in” period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

Look at trace plots, autocorrelation.

Check result with different proposal density.

If you think it’s converged, try starting from a different point and see if the result is similar.