Statistical Methods for Particle Physics (2)

CERN-FNAL
Hadron Collider Physics
Summer School
CERN, 6-15 June, 2007

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

Outline

1 Brief overview

Probability: frequentist vs. subjective (Bayesian)

Statistics: parameter estimation, hypothesis tests

2 Statistical tests for Particle Physics

multivariate methods for event selection

Wednesday

Friday

goodness-of-fit tests for discovery

3 Systematic errors

Treatment of nuisance parameters

Bayesian methods for systematics, MCMC

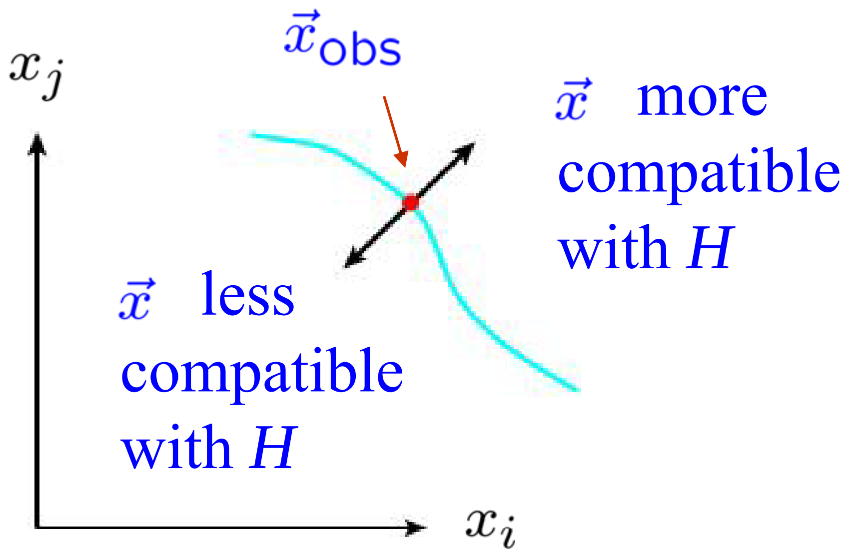
Testing goodness-of-fit

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

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

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 $\vec{x}_{\mathrm{obs}}$.
(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*.

p-value 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 = $s + b$:

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

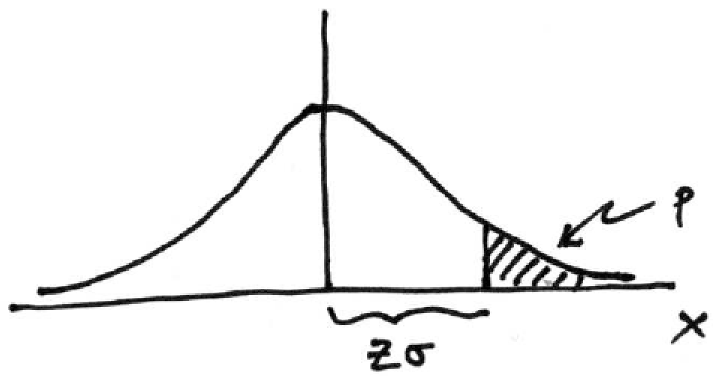
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

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.

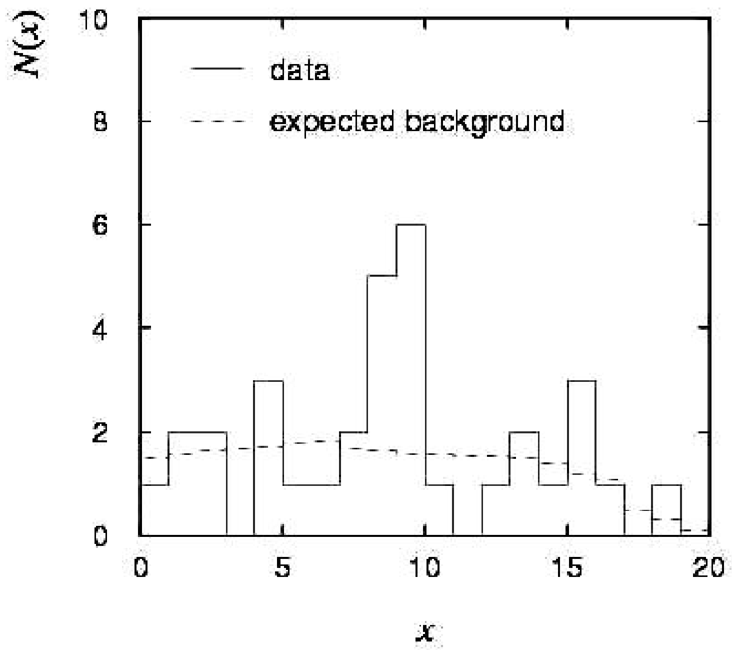


$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ **TMath::Prob**

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

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

Is the observed width consistent with the expected x resolution?

$\rightarrow$ take x window several times the expected resolution

How many bins $\times$ distributions have we looked at?

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

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

$\rightarrow$ freeze cuts, repeat analysis with new data

Should we publish????

Using shape of a distribution in a search

Suppose we want to search for a specific model (i.e. beyond the Standard Model); contains parameter θ.

Select candidate events; for each event measure some quantity x and make histogram: $\vec{n} = (n_1, \ldots, n_M)$

Expected number of entries in ith bin: $E[n_i] = s_i(\theta) + b_i$

$$E[n_i] = \underbrace{s_i(\theta)}_{\text{signal}} + \underbrace{b_i}_{\text{background}}$$

Suppose the 'no signal' hypothesis is $\theta = \theta_0$, i.e., $s(\theta_0) = 0$.

Probability is product of Poisson probabilities:

$$P(\vec{n}|\theta) = \prod_{i=1}^{M} \frac{(s_i(\theta) + b_i)^{n_i}}{n_i!} e^{-(s_i(\theta)+b_i)}$$

Testing the hypothesized θ

Construct e.g. the likelihood ratio: $t(\theta) = \frac{P(\vec{n}|\theta)}{P(\vec{n}|\theta_0)}$

Find the sampling distribution $g(t(\theta)|\theta_0)$ (e.g. use MC)
i.e. we need to know how $t(\theta)$ would be distributed if the entire experiment would be repeated under assumption of the background only hypothesis (parameter value θ_0).

p-value of θ_0 using test variable designed to be sensitive to θ:

$$p = \int_{t_{\mathrm{obs}}}^{\infty} g(t|\theta_0)\, dt$$

This gives the probability, under the assumption of background only, to see data as 'signal like' or more so, relative to what we saw.

Making a discovery / setting limits

Repeat this exercise for all θ

If we find a small p-value $\rightarrow$ discovery

Is the new signal compatible with what you were looking for?

Test hypothesized θ using $g(t(\theta)|\theta)$

If $$p = \int_{-\infty}^{t_{\rm obs}} g(t|\theta)\, dt < \alpha$$ reject θ.

here use e.g. $\alpha = 0.05$

Confidence interval at confidence level $1 - \alpha$
= set of θ values not rejected by a test of significance level α.

When to publish

HEP folklore: claim discovery when *p*-value of background only hypothesis is 2.85×10^{-7}, corresponding to significance $Z = 5$.

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

phenomenon	reasonable *p*-value for discovery
$D^0 \bar{D}^0$ mixing	~ 0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

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:

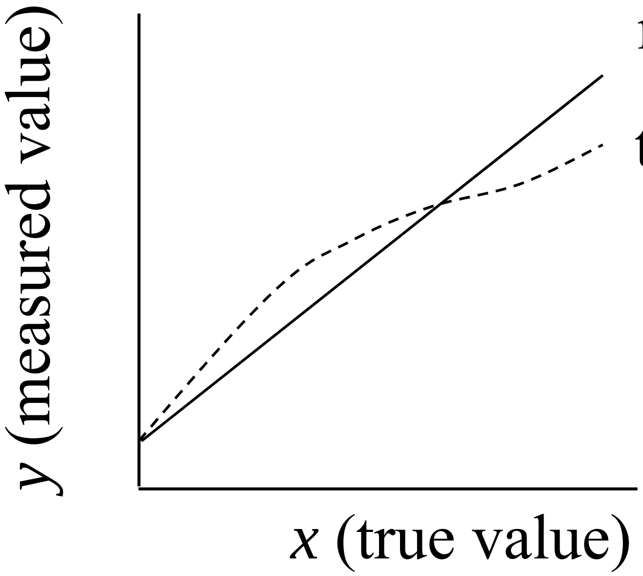
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

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, e.g., calibration constants, efficiencies, etc.

Systematic errors and nuisance parameters

Response of measurement apparatus is never modelled perfectly:



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 $\leftrightarrow$ nuisance parameters

Example: fitting a straight line

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

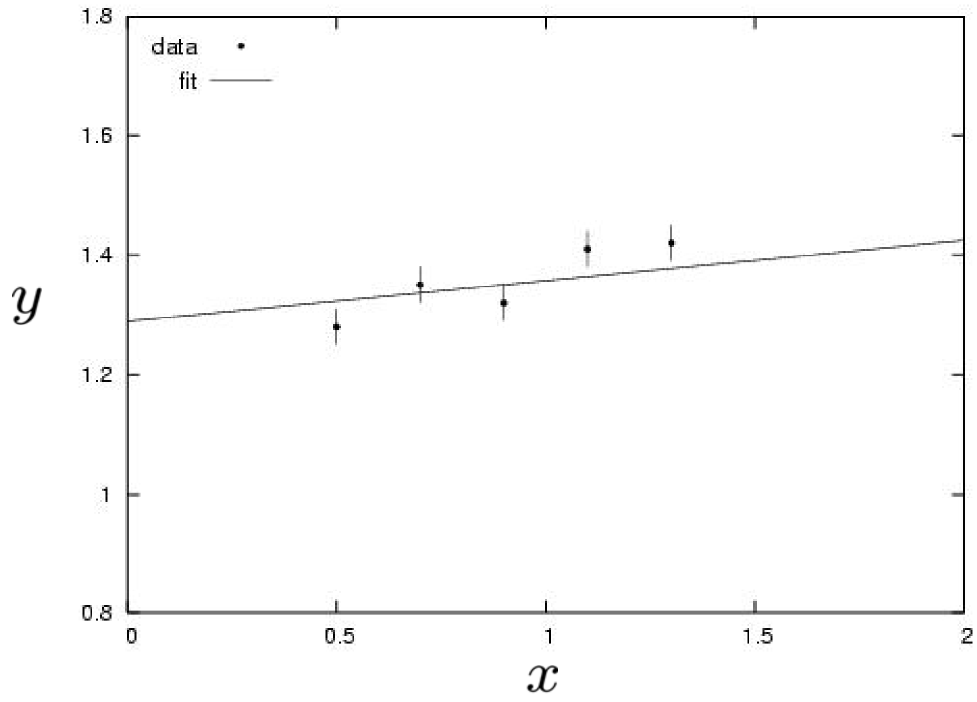
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Case #1: θ_1 known a priori

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

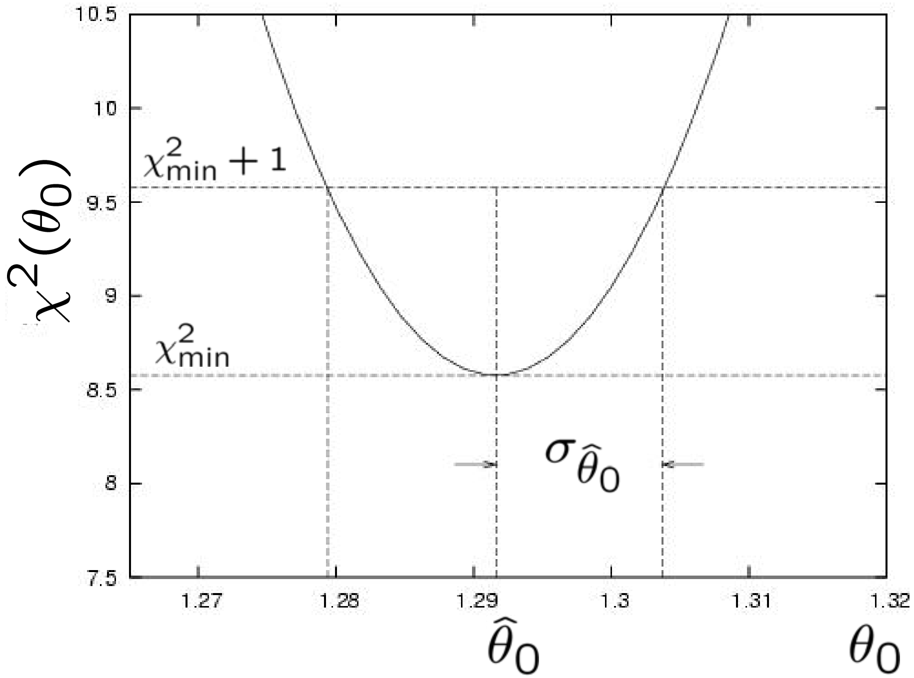
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

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

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

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



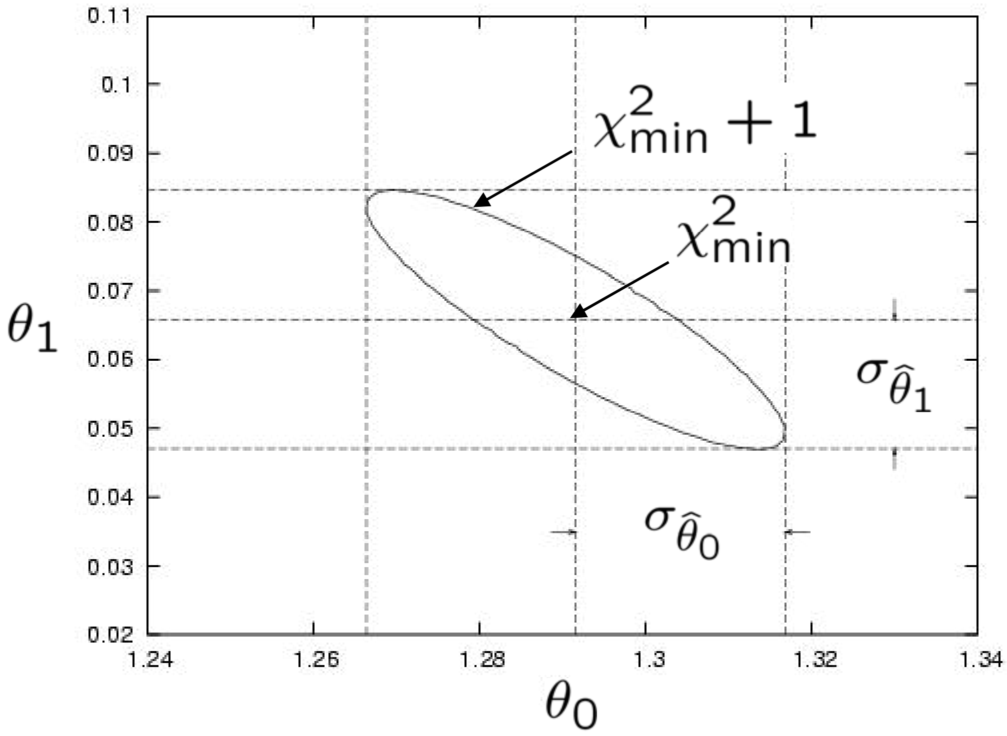
Case #2: both θ_0 and θ_1 unknown

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

Standard deviations from tangent lines to contour

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

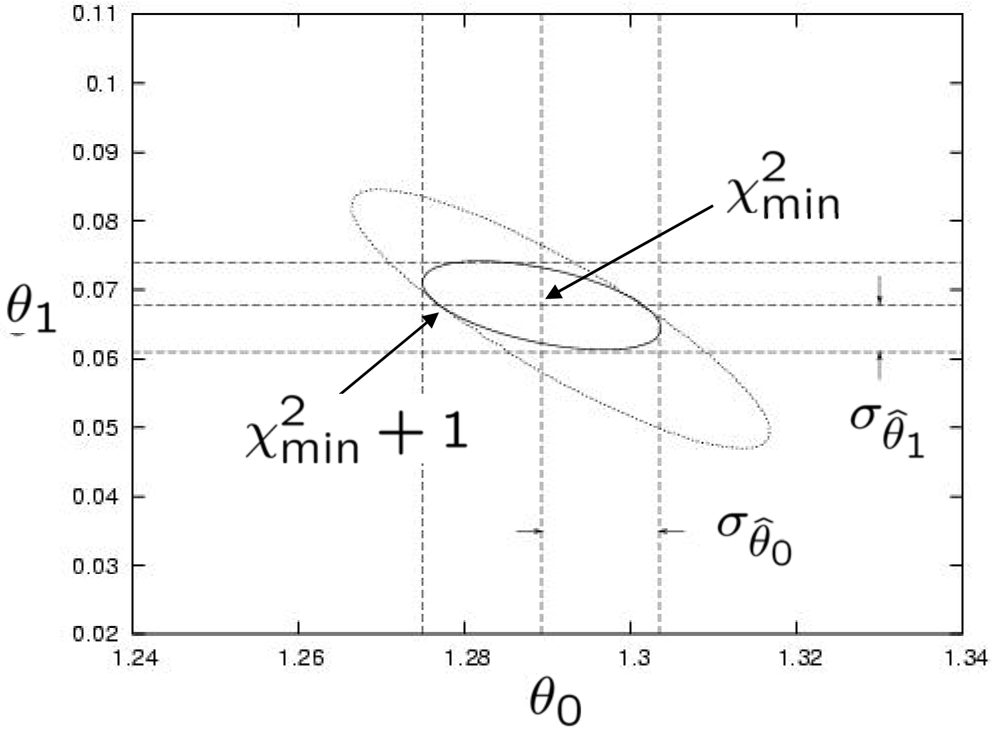
Correlation between $\hat{\theta}_0, \ \hat{\theta}_1$ causes errors to increase.



Case #3: we have 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 $\hat{\theta}_0$.



The profile likelihood

The 'tangent plane' method is a special case of using the

profile likelihood: $L'(\theta_0) = L(\theta_0, \hat{\hat{\theta}}_1)$.

$\hat{\hat{\theta}}_1$ is found by maximizing $L(\theta_0, \theta_1)$ for each θ_0.

Equivalently use $\chi^{2\prime}(\theta_0) = \chi^2(\theta_0, \hat{\hat{\theta}}_1)$.

The interval obtained from $\chi^{2\prime}(\theta_0) = \chi^{2\prime}_{\rm min} + 1$ is the same as what is obtained from the tangents to $\chi^2(\theta_0, \theta_1) = \chi^2_{\rm min} + 1$.

Well known in HEP as the 'MINOS' method in MINUIT.

Profile likelihood is one of several 'pseudo-likelihoods' used in problems with nuisance parameters.

The Bayesian approach

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

Interpret probability of θ as 'degree of belief' (subjective).

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

Our experiment has data $\boldsymbol{y}$, → likelihood function $L(\boldsymbol{y}|\theta)$.

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

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

Posterior pdf $p(\theta\,|\,\boldsymbol{y})$ contains all our knowledge about θ.

Case #4: 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 $p(\theta_0 \mid x)$:

$$p(\theta_0|\vec{y}) = \int p(\theta_0, \theta_1|\vec{y})\, d\theta_1 \ .$$

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

$$\begin{aligned} p(\theta_0|\vec{y}) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0-\hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \ \text{(same as before)} \end{aligned}$$

Ability to marginalize over nuisance parameters is an important feature of Bayesian statistics.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|\vec{y}) = \int p(\theta_0, \theta_1|\vec{y})\, d\theta_1 \ .$$

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.

Google for 'MCMC', 'Metropolis', 'Bayesian computation', ...

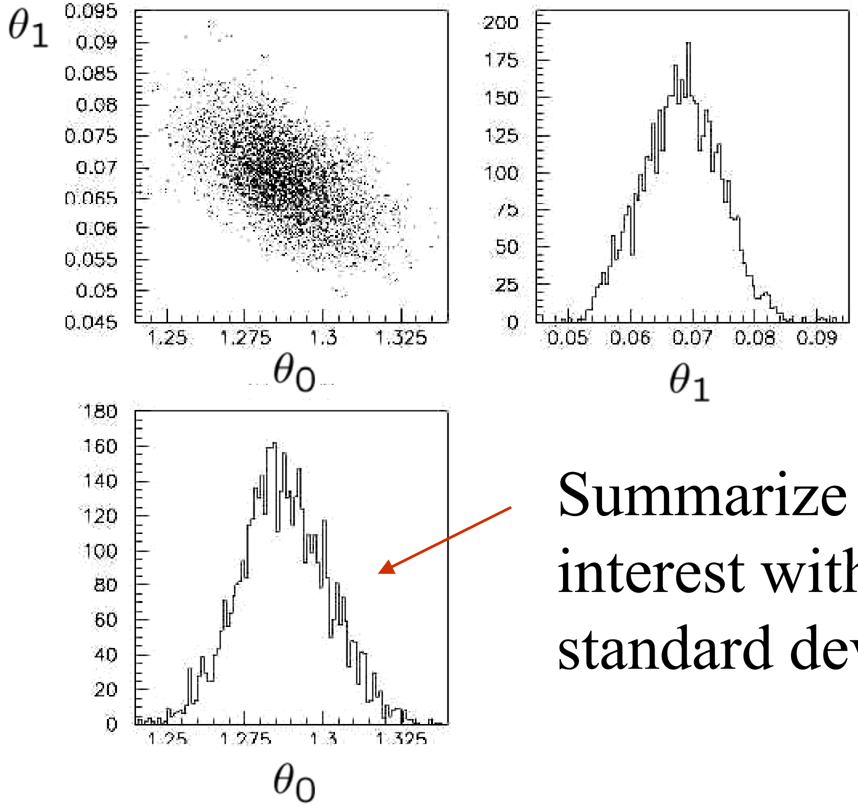
MCMC generates correlated sequence of random numbers:

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

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?)

Case #5: Bayesian method with vague prior

Suppose we don't have a previous measurement of θ_1 but rather some vague information, e.g., a theorist tells us:

$\theta_1 \geq 0$ (essentially certain);

θ_1 should have order of magnitude less than 0.1 'or so'.

Under pressure, the theorist sketches the following prior:

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

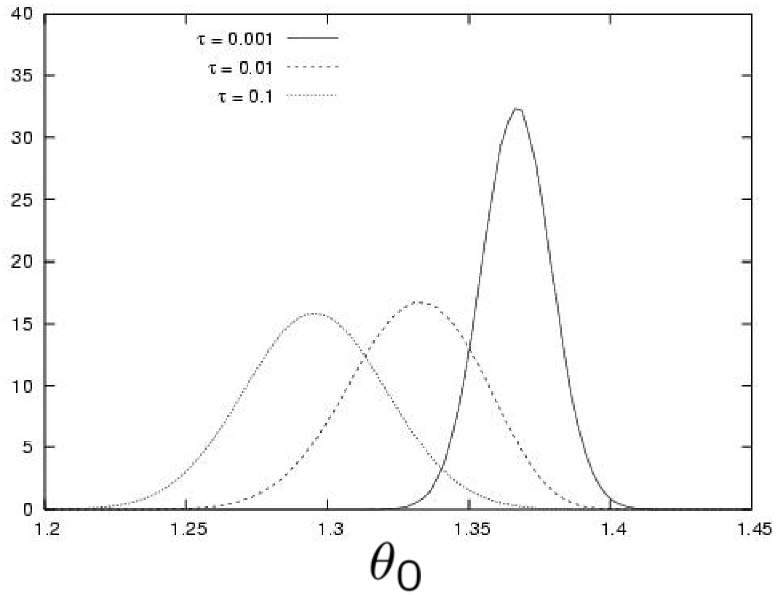
From this we will obtain posterior probabilities for θ_0 (next slide).

We do not need to get the theorist to 'commit' to this prior; final result has 'if-then' character.

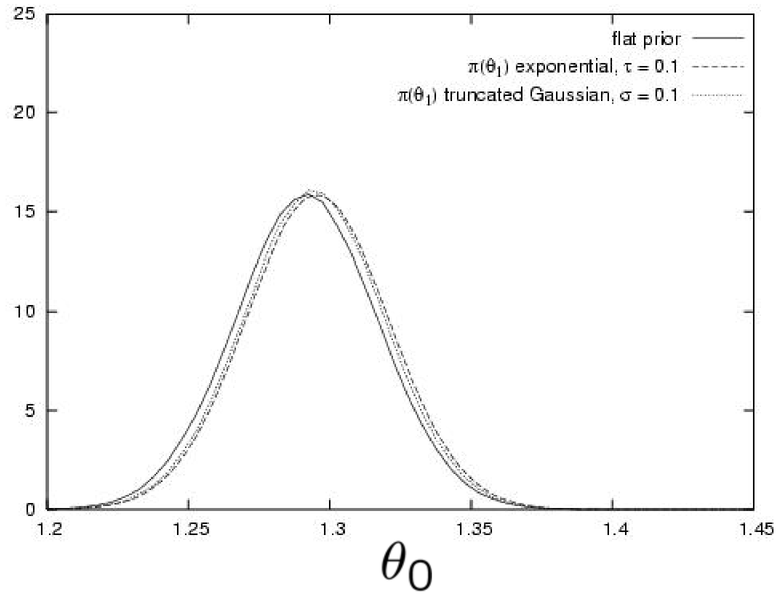
Sensitivity to prior

Vary $\pi(\theta)$ to explore how extreme your prior beliefs would have to be to justify various conclusions (sensitivity analysis).

Try exponential with different mean values...



Try different functional forms...



Wrapping up...

p-value for discovery = probability, under assumption of background only, to see data as signal-like (or more so) relative to the data you obtained.

$\neq P$(Standard Model true)!

Systematic errors $\leftrightarrow$ nuisance parameters

If constrained by measurement $\rightarrow$ profile likelihood
Other prior info $\rightarrow$ Bayesian methods

Extra slides

MCMC basics: Metropolis-Hastings algorithm

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

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

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

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, $\leftarrow$ move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ $\leftarrow$ 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.

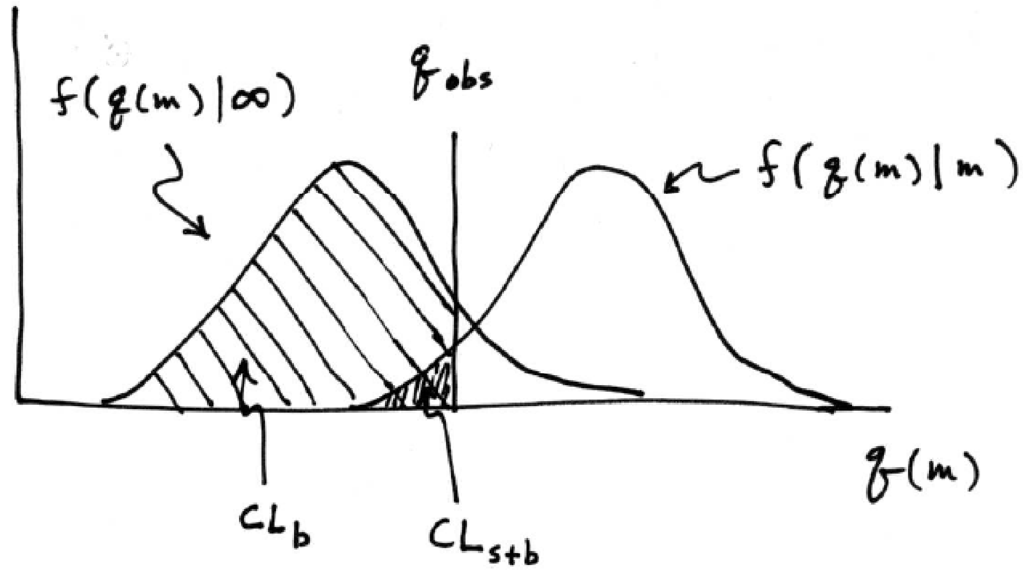
Look at trace plots, autocorrelation.

Check result with different proposal density.

If you think it's converged, try it again with 10 times more points.

LEP-style analysis: $\mathrm{CL_b}$

Same basic idea: $L(m) \to l(m) \to q(m) \to$ test of m, etc.



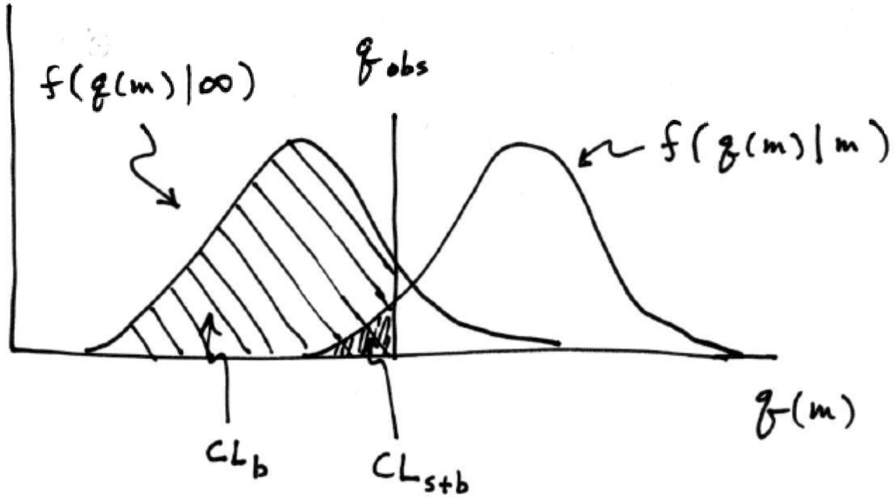
For a chosen m, find p-value of background-only hypothesis:

$$p_{\mathrm{b}} = \int_{-\infty}^{q_{\mathrm{obs}}} f(q|\infty)\, dq \equiv 1 - \mathrm{CL_b} \qquad\qquad Z = \Phi^{-1}(1 - p_{\mathrm{b}})$$

LEP-style analysis: CL_{s+b}

'Normal' way to get interval would be to reject hypothesized *m* if

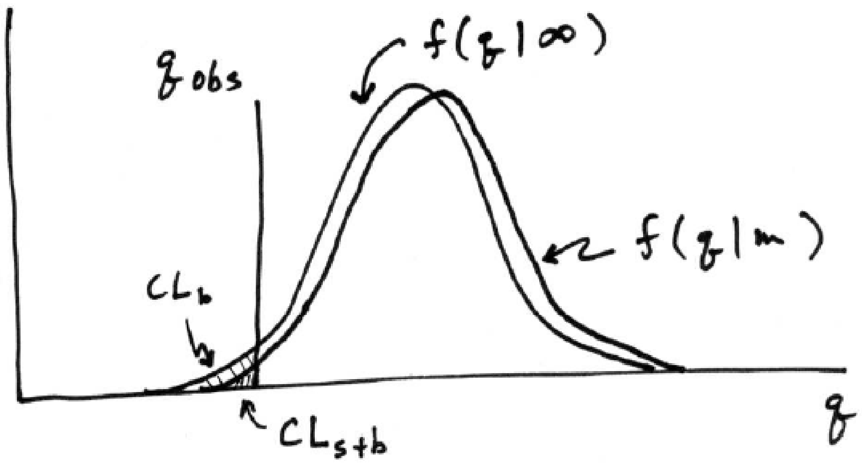
$$p-\mathrm{value}(m) = \int_{-\infty}^{q_{\mathrm{obs}}} f(q|m)\, dq \equiv \mathrm{CL}_{\mathrm{s+b}} < \alpha \;.$$



By construction this interval will cover the true value of m with probability $1-\alpha$.

LEP-style analysis: CL_s

The problem with the CL_{s+b} method is that for high m, the distribution of q approaches that of the background-only hypothesis:



So a low fluctuation in the number of background events can give $\mathrm{CL}_{s+b} < \alpha$

This rejects a high m value even though we are not sensitive to Higgs production with that mass; the reason was a low fluctuation in the background.

$\mathrm{CL_s}$

A solution is to define: $$\mathrm{CL_s} = \frac{\mathrm{CL_{s+b}}}{\mathrm{CL_b}}$$

and reject the hypothesized *m* if: $$\mathrm{CL_s} \leq \alpha \ .$$

Since $\mathrm{CL_b} \leq 1$, one has $\mathrm{CL_s} \geq \mathrm{CL_{s+b}}$.

So the $\mathrm{CL_s}$ intervals 'over-cover'; they are conservative.