Statistical Methods in HEP, in particular... Nuisance parameters and systematic uncertainties

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

RHUL HEP Seminar
22 March, 2006

A rehash my talk at the IoP Half Day Meeting on Statistics in HEP
University of Manchester, 16 November, 2005

Itself a rehash of PHYSTAT 2005, Oxford, September 2005

Vague outline

I. Nuisance parameters and systematic uncertainty

II. Parameter measurement
- Frequentist
- Bayesian

III. Estimating intervals (setting limits)
- Frequentist
- Bayesian

IV. Comment on the D0 result on B_s mixing

V. Conclusions

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:

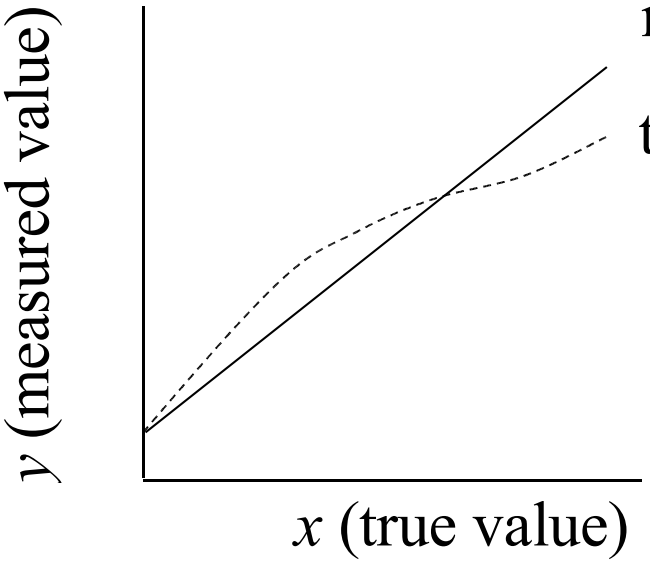
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling of measurement apparatus.

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

Nuisance parameters

Suppose the outcome of the experiment is some set of data values x (here shorthand for e.g. $x_1, \ldots, x_n$).

We want to determine a parameter θ,
(could be a vector of parameters $\theta_1, \ldots, \theta_n$).

The probability law for the data x depends on θ :

$L(x|\,\theta)$ (the likelihood function)

E.g. maximize L to find estimator $\hat{\theta}$.

Now suppose, however, that the vector of parameters: $\theta = (\psi, \lambda)$

contains some that are of interest, $\psi_1, \ldots, \psi_n$,
and others that are not of interest: $\lambda_1, \ldots, \lambda_m$.
Symbolically:

The $\lambda_1, \ldots, \lambda_m$ are called nuisance parameters.

Example #1: fitting a straight line

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

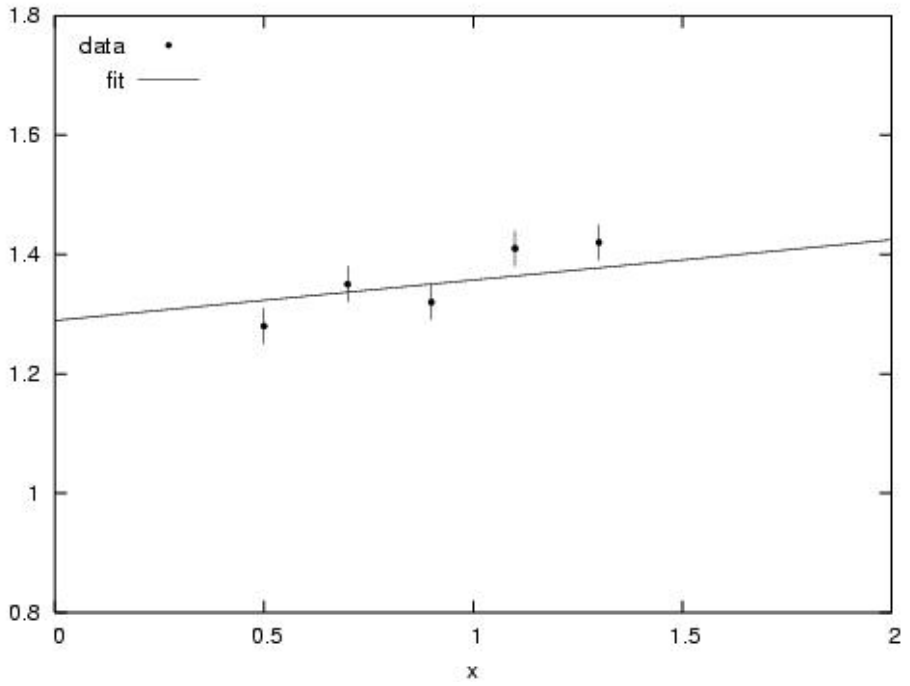
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Case #1: θ_1 known a priori

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

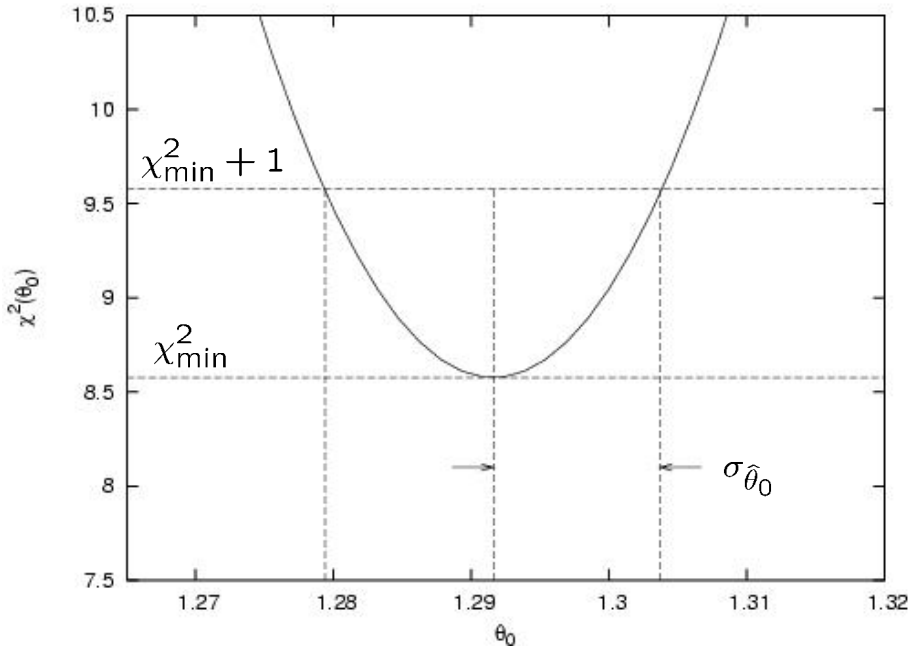
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

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

Come up one unit from χ^2_{min}

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



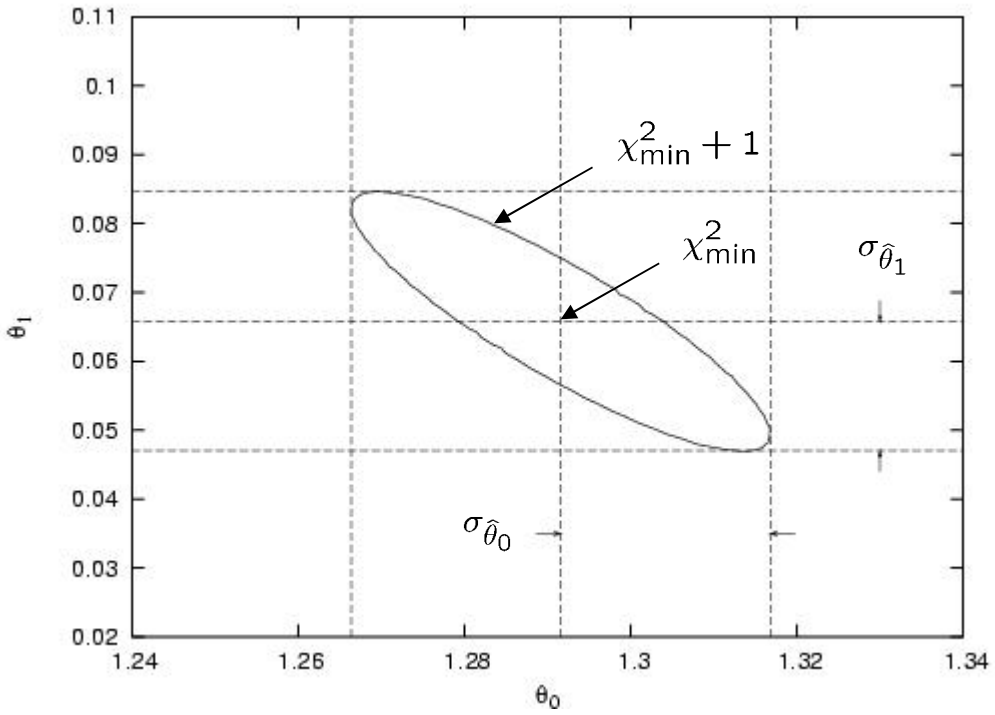
Case #2: both θ_0 and θ_1 unknown

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

Standard deviations from tangent lines to contour

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

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

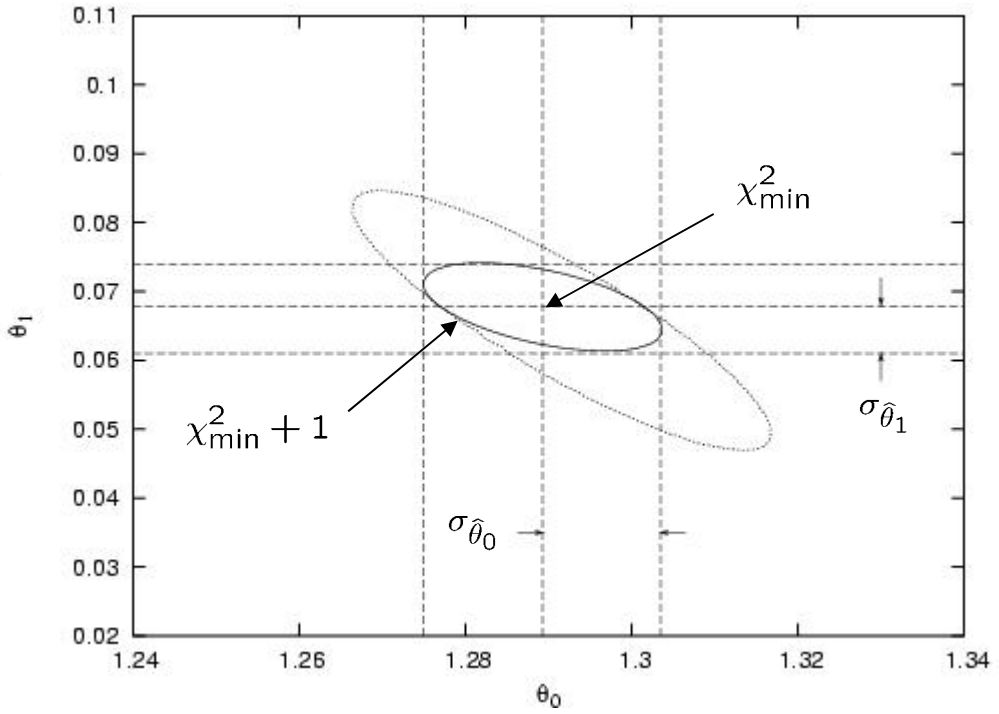


Case #3: we have a measurement t_1 of θ_1

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

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

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

Profile likelihood is one of several 'pseudo-likelihoods' used in problems with nuisance parameters. See e.g. talk by Rolke at PHYSTAT05.

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 x, → 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 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 Θ 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

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

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

Digression: marginalization with MCMC

Bayesian computations involve integrals like

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

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.

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.

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 $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathsf{Uniform}[0, 1]$

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.

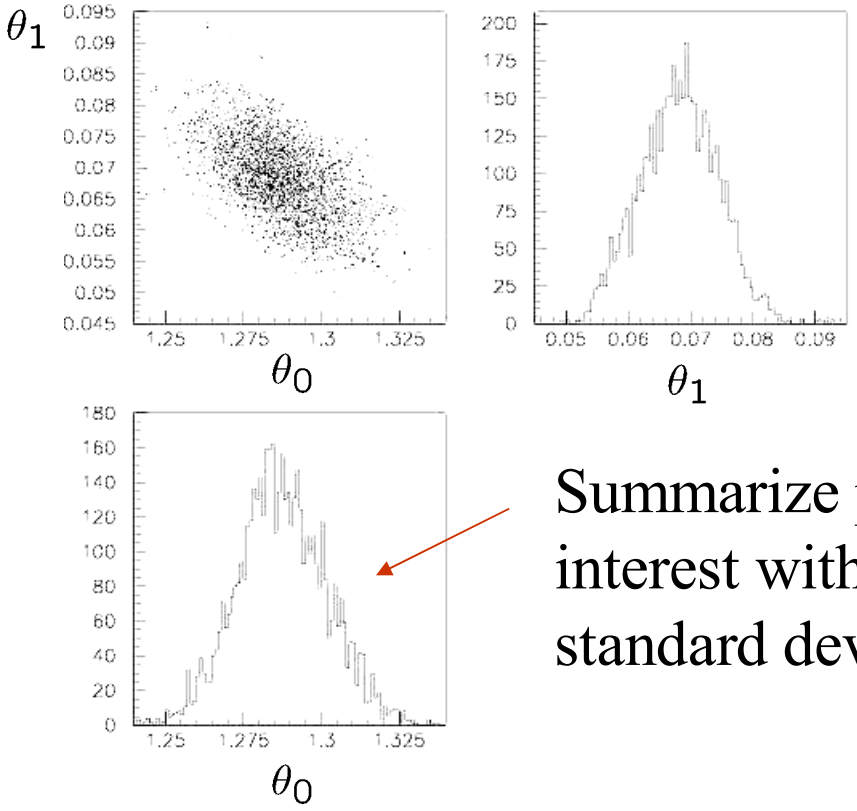
Look at trace plots, autocorrelation.

Check result with different proposal density.

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

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:

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

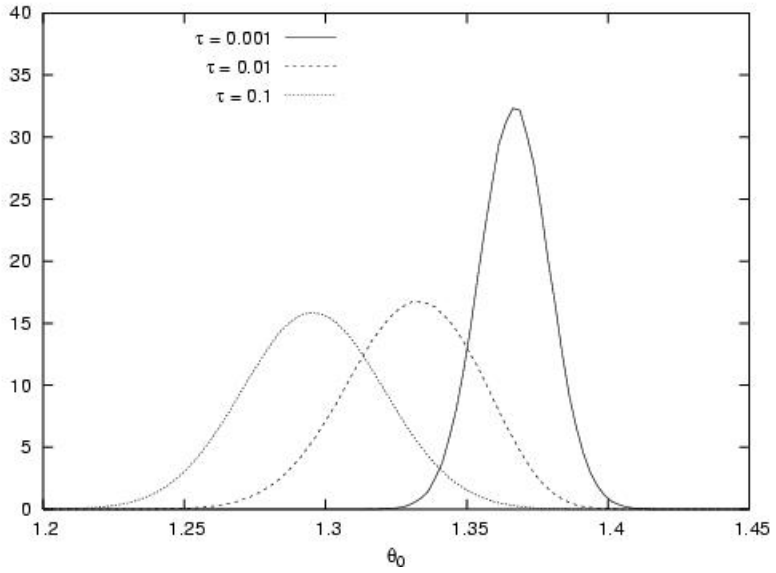
From this we will obtain posterior probabilities for θ_0 (next slide).

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

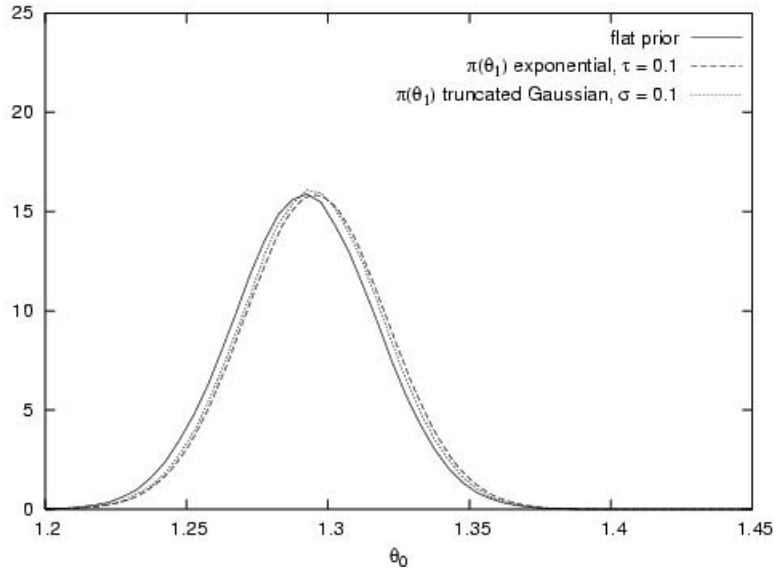
Sensitivity to prior

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...



Example #2: 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

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

Sometimes b known, other times it is in some way uncertain.

Goal: measure or place limits on s, taking into consideration the uncertainty in b.

Widely discussed in HEP community, see e.g. proceedings of PHYSTAT meetings, Durham, Fermilab, CERN workshops...

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.
>
> (Because of discrete data, need inequality here.)
>
> 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 $\geq 1 - \gamma$.

Equivalent to Neyman confidence belt construction.

Setting limits: 'classical method'

E.g. for upper limit on s, take critical region to be low values of n, limit s_{up} at confidence level $1 - \beta$ thus found from

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

Similarly for lower limit at confidence level $1 - \alpha$,

$$\alpha = P(n \geq n_{\text{obs}}; s_{\text{lo}}, b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{(s_{\text{lo}} + b)^n}{n!} e^{-(s_{\text{lo}}+b)} .$$

Sometimes choose $\alpha = \beta = \gamma/2$ $\rightarrow$ central confidence interval.

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 $0 \leq l(s) \leq 1$.

Critical region defined by low values of likelihood ratio.

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.

Nuisance parameters and limits

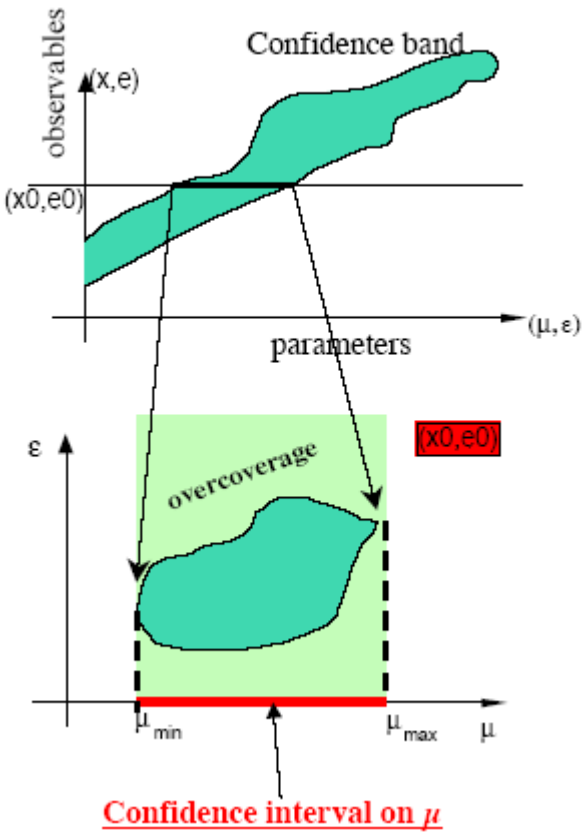
In general we don't know the background b perfectly.

Suppose we have a measurement of b, e.g., $b_{\text{meas}} \sim N(b, \sigma_b)$

So the data are really: n events and the value b_{meas}.

In principle the confidence interval recipe can be generalized to two measurements and two parameters.

Difficult and rarely attempted, but see e.g. talk by G. Punzi at PHYSTAT05.



G. Punzi, PHYSTAT05

Bayesian limits with uncertainty on b

Uncertainty on b goes into the prior, e.g.,

$$\pi(s,b) = \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)}$$

$$\pi_s(s) = \text{const}, \quad \sim 1/s, \ldots \quad \text{? (see R. Barlow talk)}$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over b, then use $p(s|n)$ to find intervals for s with any desired probability content.

Controversial part here is prior for signal $\pi_s(s)$
(treatment of nuisance parameters is easy).

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 CL = 1 – β: $\beta = P(n \leq n_{\text{obs}}; s_{\text{up}})$

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 (FHC[2]), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

Interval from inverting profile LR test

Suppose we have a measurement b_{meas} of b.

Build the likelihood ratio test with profile likelihood:

$$l(s) = \frac{L(n, b_{\text{meas}}|s, \hat{\hat{b}})}{L(n, b_{\text{meas}}|\hat{s}, \hat{b})}$$

and use this to construct confidence intervals.

See PHYSTAT05 talks by Cranmer, Feldman, Cousins, Reid.

Comment on B_s mixing from D0

Last week D0 announced the discovery of B_s mixing:

Moriond talk by Brendan Casey, also hep-ex/0603029

Produce a B_q meson at time t=0; there is a time dependent probability for it to decay as an anti-B_q (q = d or s):

$$P(\overline{B}_q(t)|B_q(0)) \approx \frac{\Gamma}{2} e^{-\Gamma t}(1 - \cos \Delta m_q t) \qquad \Delta m_q \sim |V_{tb}^* V_{tq}|^2$$

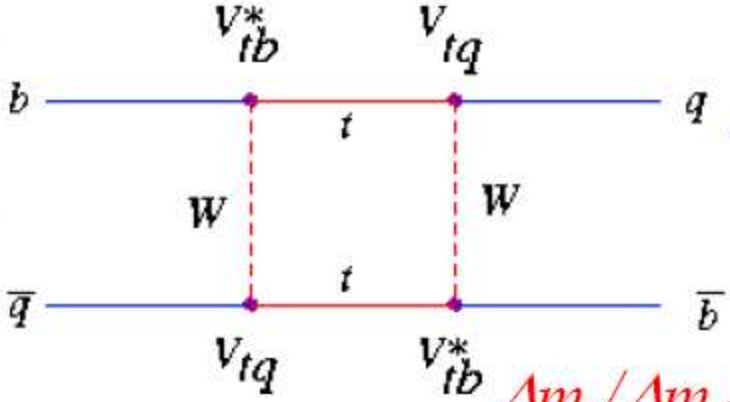
$|V_{ts}| \gg |V_{td}|$ and so B_s oscillates quickly compared to decay rate

Sought but not seen at LEP;
early on predicted to be visible at Tevatron

Here are some of Casey's slides with commentary...

Why B_s Mixing?

#1: It's a QCD thing

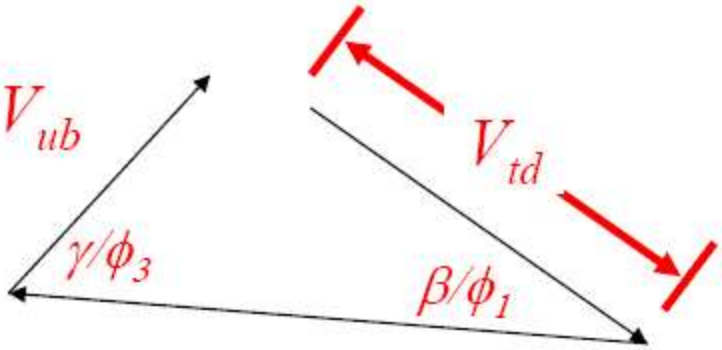


$\Delta m_d \sim$ (lots of QCD)x$|V_{td}|$

$\Delta m_s \sim$ (lots of QCD)x$|V_{ts}|$

$\Delta m_s / \Delta m_d \sim$ (much less QCD)x$|V_{ts}|/|V_{td}|$

#2: why V_{td}?
New sources of CPV



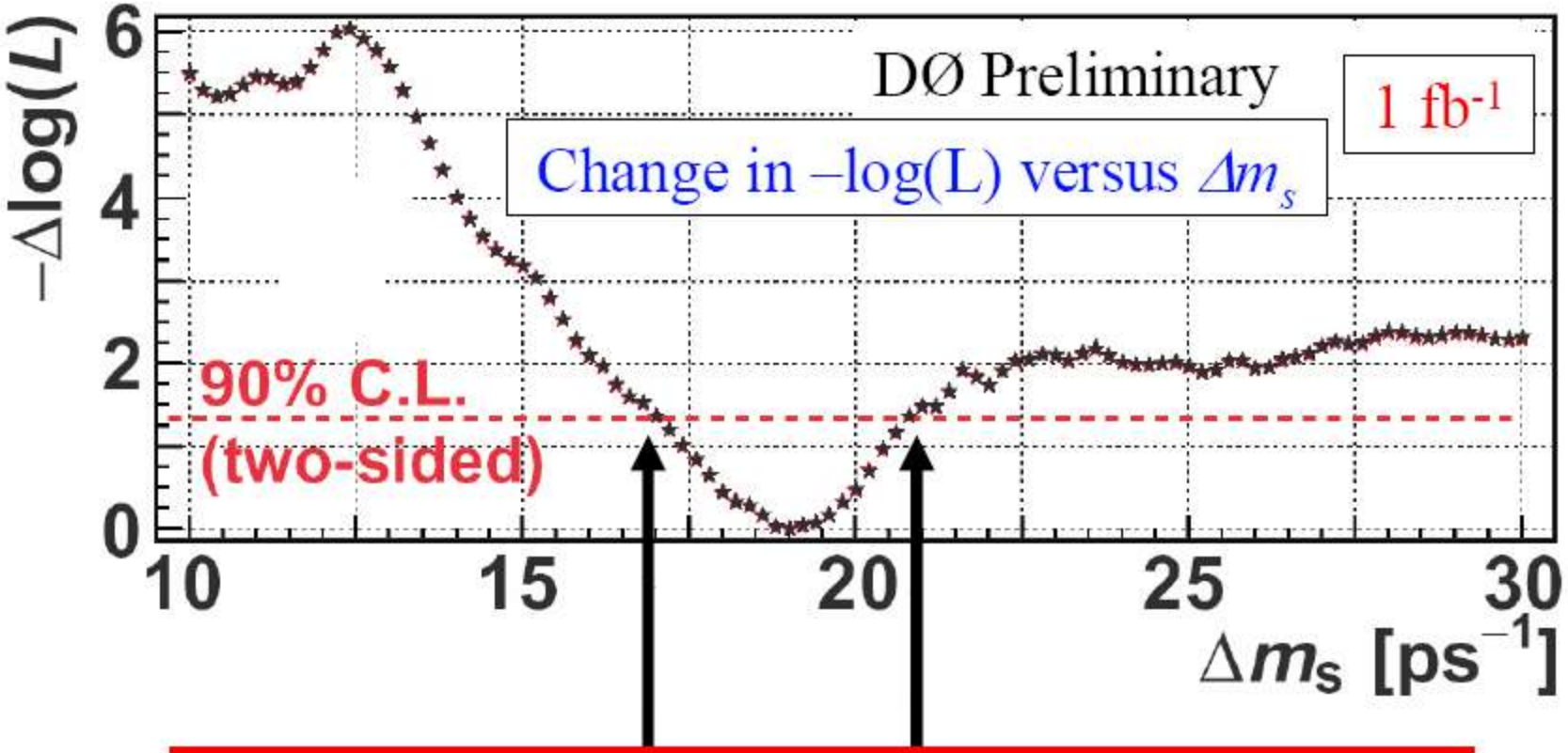
2 fundamental CKM relations:

$|V_{ub}|$ constrains $\beta/\phi_1 = Arg(V_{td})$ ✓

$|V_{td}|$ constrains $\gamma/\phi_3 = Arg(V_{ub})$?

Disagreement = new source of CPV

Switch Back to Likelihood Fit



$17 < \Delta m_s < 21\ \text{ps}^{-1}$ @ 90% CL

Maximum likelihood at $\Delta m_s = 19\ \text{ps}^{-1}$

First time Δm_s bounded on both sides!!!!!

Confidence interval from likelihood function

In the large sample limit it can be shown for ML estimators:

$\hat{\vec{\theta}} \sim N(\vec{\theta}, V)$ (n-dimensional Gaussian, covariance V)

$$L(\vec{\theta}) = L_{\max} \exp\left[-\frac{1}{2} Q(\hat{\vec{\theta}}, \vec{\theta})\right], \quad Q(\hat{\vec{\theta}}, \vec{\theta}) = (\hat{\vec{\theta}} - \vec{\theta})^T V^{-1} (\hat{\vec{\theta}} - \vec{\theta})$$

$Q(\hat{\vec{\theta}}, \vec{\theta}) = Q_\gamma$ defines a hyper-ellipsoidal confidence region,

$$P(\text{ellipsoid covers true } \vec{\theta}) = P(Q(\hat{\vec{\theta}}, \vec{\theta}) \leq Q_\gamma)$$

If $\hat{\vec{\theta}} \sim N(\vec{\theta}, V)$ then $Q(\hat{\vec{\theta}}, \vec{\theta}) \sim \text{Chi-square}(n)$

$$\text{coverage probability} \equiv 1 - \gamma = \int_0^{Q_\gamma} f_{\chi^2}(z; n)\, dz = F_{\chi^2}(Q_\gamma; n)$$

Approximate confidence regions from $L(\theta)$

So the recipe to find the confidence region with CL = $1-\gamma$ is:

$$\ln L(\vec{\theta}) = \ln L_{\max} - \frac{Q_\gamma}{2} \quad \text{or} \quad \chi^2(\vec{\theta}) = \chi^2_{\min} + Q_\gamma$$

where $Q_\gamma = F^{-1}_{\chi^2}(1-\gamma; n)$

Q_γ	$1-\gamma$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
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

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
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

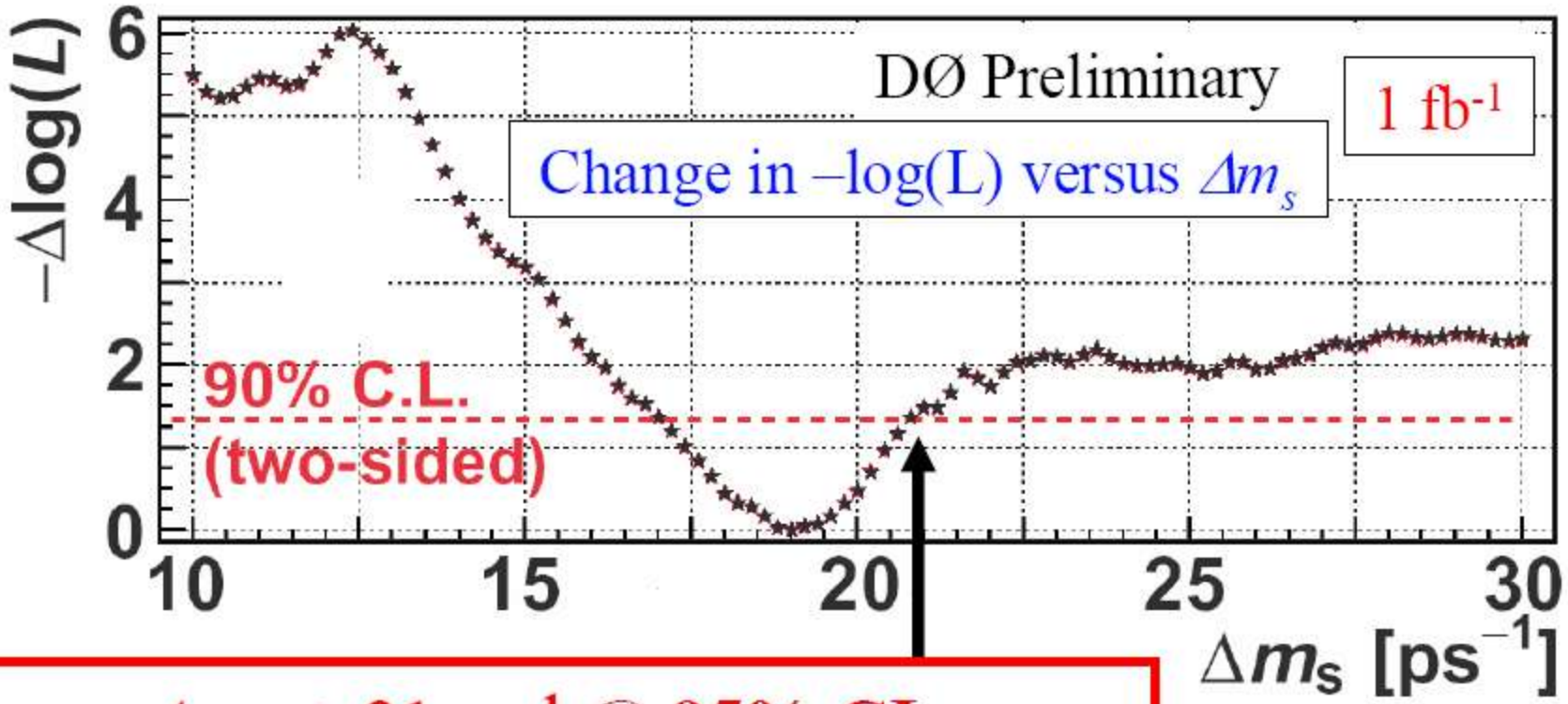
For finite samples, these are approximate confidence regions.

Coverage probability not guaranteed to be equal to $1-\gamma$;

no simple theorem to say by how far off it will be (use MC).

Remember here the interval is random, not the parameter.

Switch Back to Likelihood Fit



$\Delta m_s < 21\ ps^{-1}$ @ 95% CL

For one sided limit

5% probability for Δm_s to lie anywhere between 21 ps^{-1} and infinity

Significance Tests

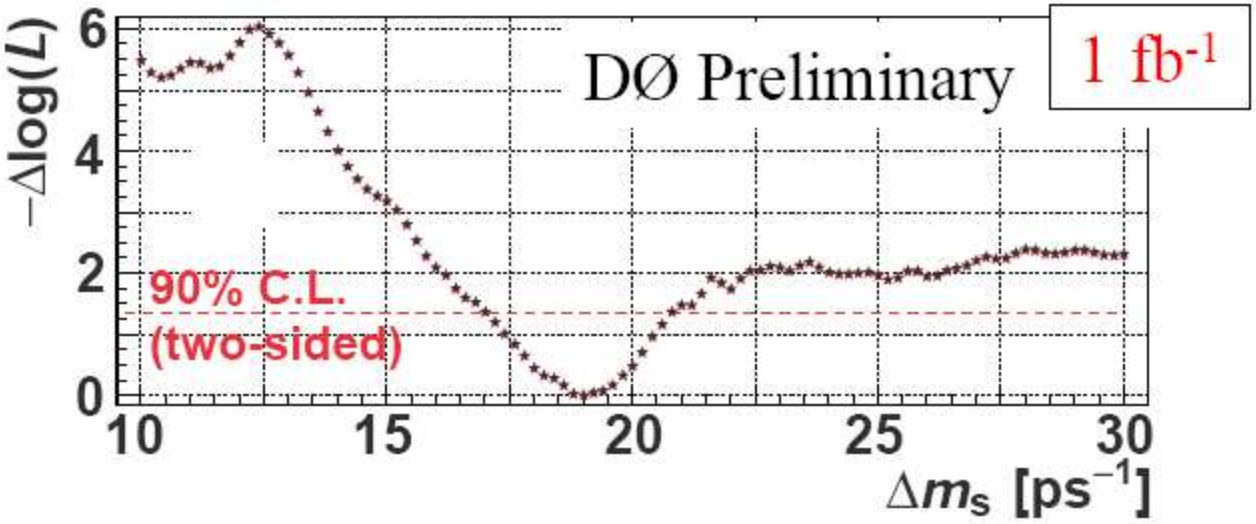
Probability for this measurement to lie in the range

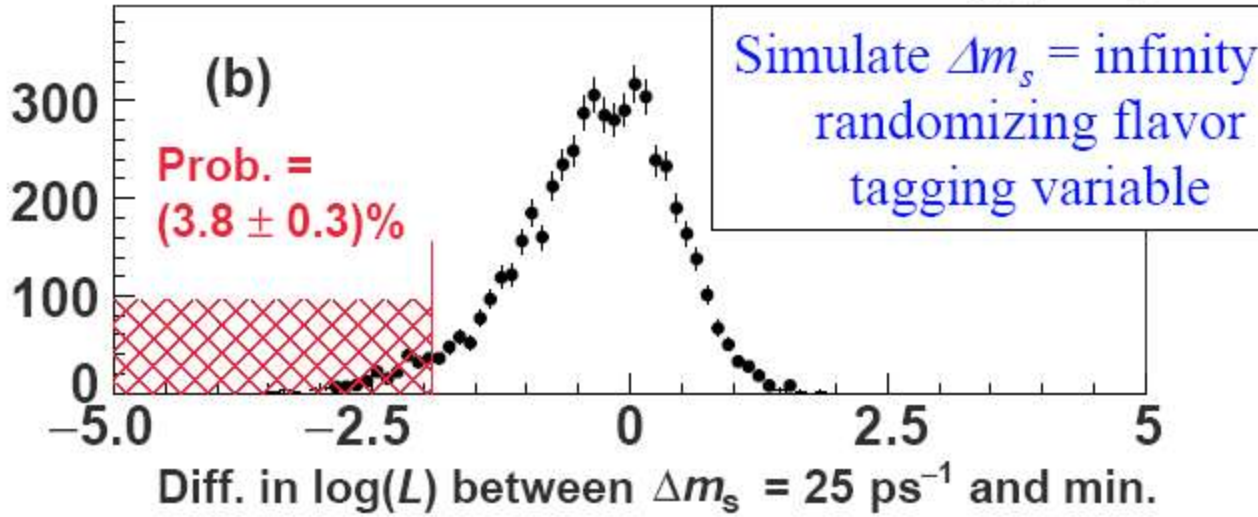
$16 < \Delta m_s < 22 ps^{-1}$

Given a true value:

Δm_s = infinity: 3.8%

$\Delta m_s = 19\ ps^{-1}$: 15%





Simulate Δm_s = infinity by randomizing flavor tagging variable

Consistent results from parameterized MC studies

Upper limit from test of hypothesized Δm_s

Base test on likelihood ratio (here $\theta = \Delta m_s$): $\quad l(\theta) = \dfrac{L(\theta)}{L(\hat{\theta})}$

Observed value is l_{obs}, sampling distribution is $g(l;\theta)$ (from MC)

θ is excluded at CL=1$-\gamma$ if $\displaystyle\int_0^{l_{\text{obs}}} g(l;\theta)\, dl \leq \gamma$

D0 shows the distribution of ln l for $\Delta m_s = 25$ ps^{-1}

$$\ln l_{\text{obs}} \approx -2\,,$$

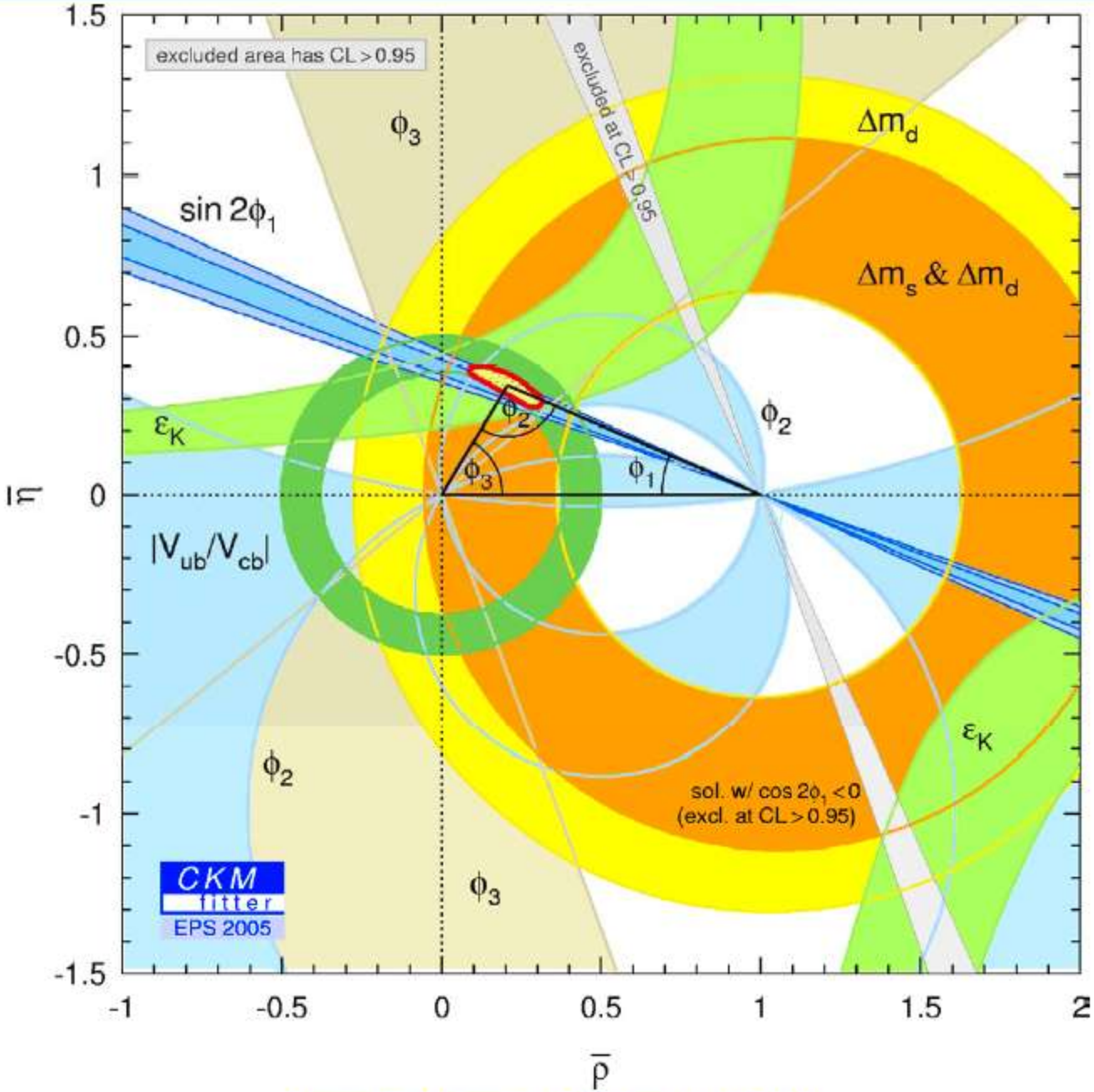
$$P(\ln l < -2; \Delta m_s = 25\,\text{ps}^{-1}) = 0.038$$

equivalent to 2.1σ effect

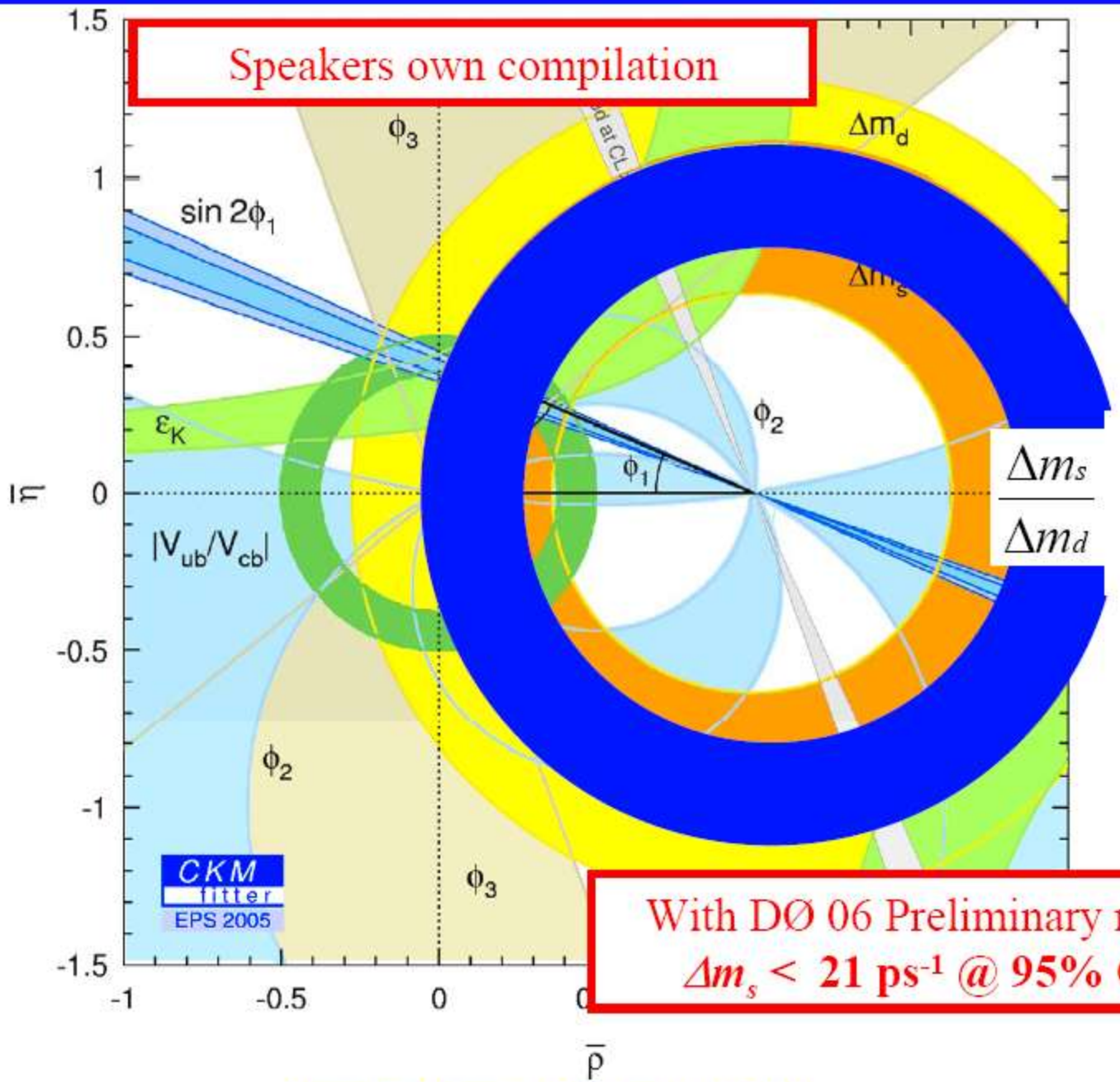
$$P(\ln l < \ln l_{\text{obs}}; \Delta m_s = 22\,\text{ps}^{-1}) = 0.05$$

95% CL upper limit

New CKM Triangle



New CKM Triangle



$\frac{\Delta m_s}{\Delta m_d}$

95% CL upper bound

With DØ 06 Preliminary result

$\Delta m_s < 21\ \text{ps}^{-1}$ @ 95% CL

Wrapping up

I've shown a few ways of treating nuisance parameters in two examples (fitting line, Poisson mean with background).

No guarantee this will bear any relation to the problem you need to solve...

At recent PHYSTAT meetings the statisticians have encouraged physicists to:

learn Bayesian methods,

don't get too fixated on coverage,

try to see statistics as a 'way of thinking' rather than a collection of recipes.

I tend to prefer the Bayesian methods for systematics but still a very open area of discussion.