Likelihoods

1) Introduction .
2) Do's & Dont's

Louis Lyons and Lorenzo Moneta

Imperial College & Oxford

CERN

CERN Academic Training Course

Nov 2016

Topics
(mainly Parameter Determination)

What it is

How it works: Resonance

Uncertainty estimates

Detailed example: Lifetime

Several Parameters

Extended maximum $\mathcal{L}$

Do's and Dont's with $\mathcal{L}$

Simple example: Angular distribution

Start with pdf = Prob density fn for data, given param values:

$$y = N(1 + \beta \cos^2\theta)$$

$y_i = N(1 + \beta \cos^2\theta_i)$

= probability density of observing θ_i, given β

$\mathcal{L}(\beta) = \Pi\, y_i$

= probability density of observing the data set y_i, given β

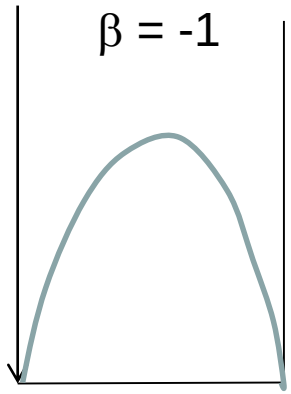
Best estimate of β is that which maximises $\mathcal{L}$

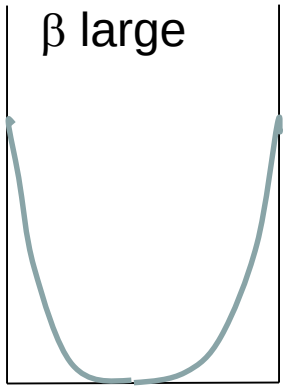
Values of β for which $\mathcal{L}$ is very small are ruled out

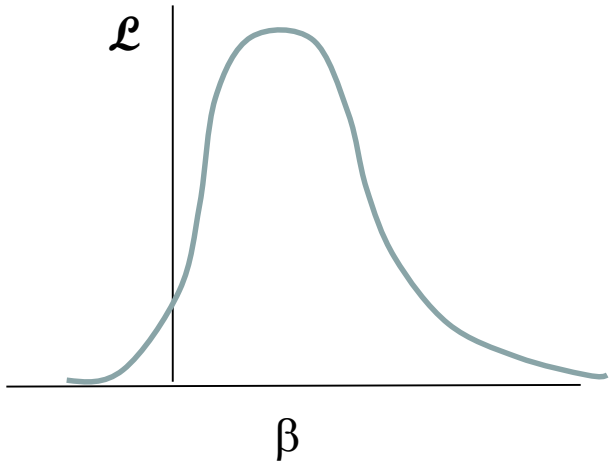
Precision of estimate for β comes from width of $\mathcal{L}$ distribution

CRUCIAL to normalise y $N = 1/\{2(1 + \beta/3)\}$

(Information about parameter β comes from **shape** of exptl distribution of $\cos\theta$)



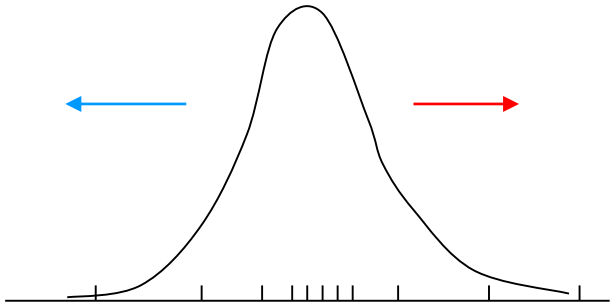




How it works: Resonance

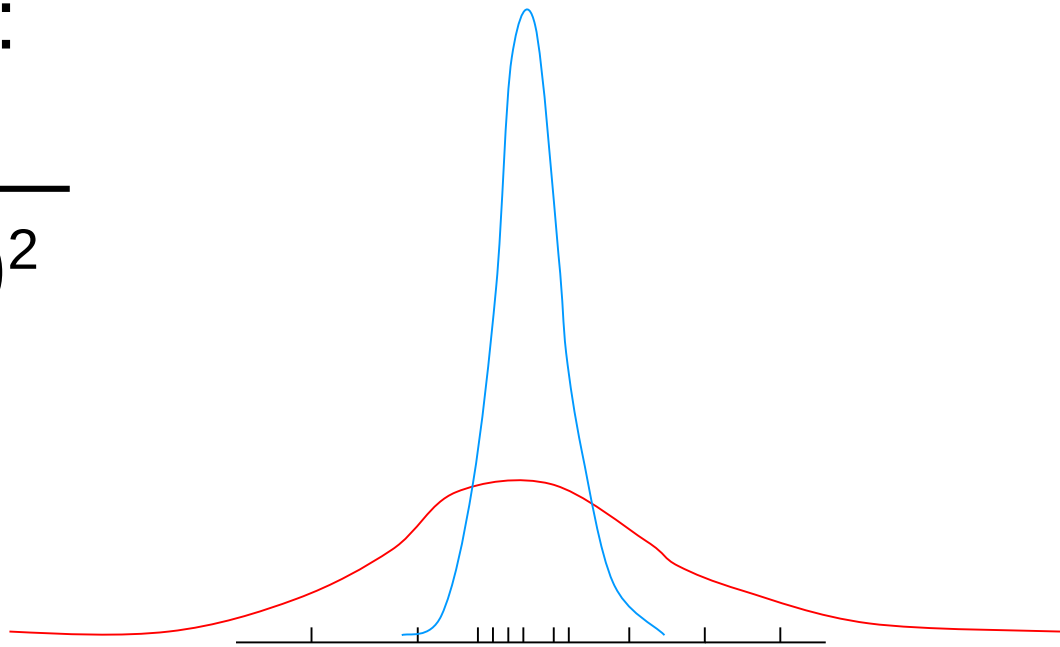
First write down pdf:

$$y \sim \frac{\Gamma/2}{(m-M_0)^2 + (\Gamma/2)^2}$$



m →

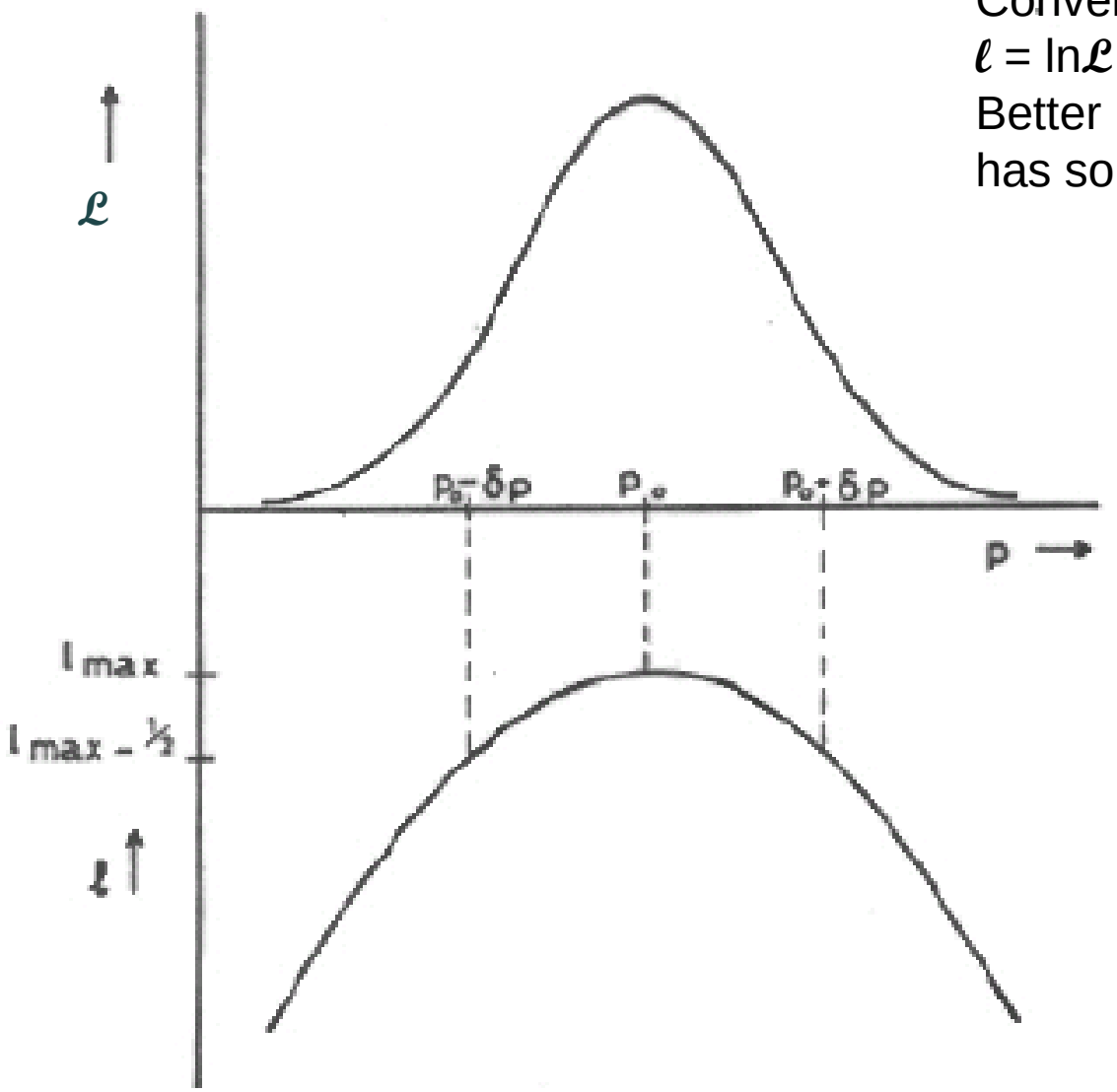
Vary M_0



m →

Vary Γ

N.B. Can make use of individual events



Conventional to consider
$\ell = \ln\mathcal{L} = \Sigma \ln y_i$
Better numerically, and has some nice properties

Maximum likelihood uncertainty

Range of likely values of param μ from width of $\mathcal{L}$ or l dists.

If $\mathcal{L}(\mu)$ is Gaussian, following definitions of σ are equivalent:

1) RMS of $\mathcal{L}(\mu)$

2) $1/\sqrt{(-d^2 \ln \mathcal{L} / d\mu^2)}$ (Mnemonic)

3) $\ln(\mathcal{L}(\mu_0 \pm \sigma) = \ln(\mathcal{L}(\mu_0)) - 1/2$

If $\mathcal{L}(\mu)$ is non-Gaussian, these are no longer the same

~~"Procedure 3) above still gives interval that contains the true value of parameter μ with 68% probability"~~

Return to 'Coverage' later

Uncertainties from 3) usually asymmetric, and asym uncertainties are messy. So choose param sensibly

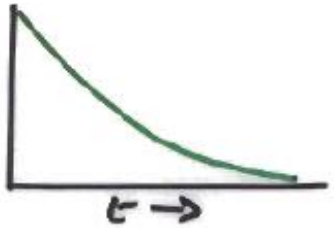
e.g 1/p rather than p; τ or λ

Lifetime Determination

Realistic analyses are more complicated than this

$$\frac{dn}{dt} = \frac{1}{\tau} e^{-t/\tau}$$

↑ NORMALISATION



Observe $t_1, t_2 \ldots\ldots t_N$

Use pdf to construct

$$\mathcal{L} = \Pi \left(\frac{dn}{dt}\right)_i = \Pi \frac{1}{\tau} e^{-t_i/\tau}$$

$$\therefore \ell = \sum_i \left(-t_i/\tau - \ln \tau\right)$$

$$\frac{\partial \ell}{\partial \tau} = \sum \left(+t_i/\tau^2 - \frac{1}{\tau}\right) = 0 = \frac{\Sigma t_i}{\tau^2} - \frac{N}{\tau}$$

$$\Rightarrow \tau = \Sigma t_i / N = \overline{t_i}$$

"Obvious"

$$\frac{\partial^2 \ell}{\partial \tau^2} = -\sum \frac{2t_i}{\tau^3} + \sum \frac{1}{\tau^2} = -2\frac{N}{\tau^2} + \frac{N}{\tau^2} = -\frac{N}{\tau^2}$$

$$\Rightarrow \sigma_\tau = 1/\sqrt{-\frac{\partial^2 \ell}{\partial \tau^2}} = \tau/\sqrt{N}$$

N.B. 1) Usual $1/\sqrt{N}$ behaviour

2) $\sigma_\tau \propto \tau_{est}$

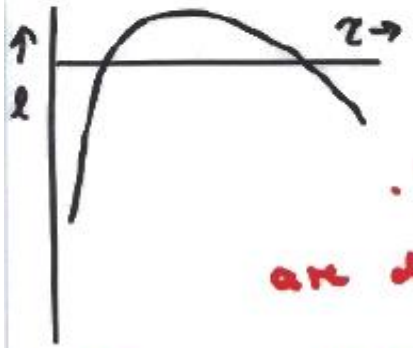
BEWARE FOR AVERAGING RESULTS

$\ln \tau - \ln \tau_{max}$ = Universal Fn of τ/τ_{max}

$$\ell(\tau) = \sum -t_i/\tau - N \ln \tau$$

$$\ell(\tau) - \ell(\tau_{max}) = -N\tau_{max}/\tau - N \ln \tau + N + N \ln \tau_{max}$$

$$= N\left[1 + \ln(\tau_{max}/\tau) - \tau_{max}/\tau\right]$$



∴ For given N, σ_+ & σ_- are defined ($\sim \frac{\tau_{max}}{\sqrt{N}}$ as $N \to \infty$)

For small N, $\sigma_+ > \sigma_-$

— " —

$$\ell(\tau_{max}) = -N(1 + \ln \bar{t})$$

N.B. $\ell(\tau_{max})$ depends only on $\bar{t}$, but not on distribution of t_i

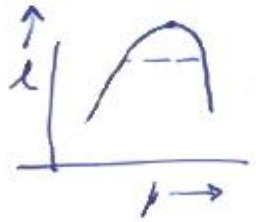
Relevant for whether ℓ_{max} is useful for testing goodness of fit

Several Parameters

1 param β

β from $\frac{\partial \ell}{\partial \beta} = 0$

$\sigma_\beta^2 = 1/\left(-\frac{\partial^2 \ell}{\partial \beta^2}\right)$



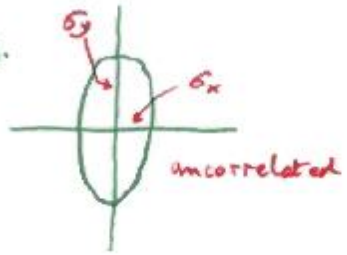
Many dimensions : $\ell(\beta_1, \beta_2, \beta_3 \ldots)$

$\beta_1, \beta_2, \beta_3 \ldots$ from $\frac{\partial \ell}{\partial \beta_i} = 0$

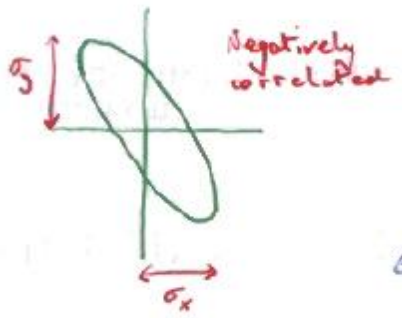
For errors, define $H_{ij} = -\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j}$ = Inverse Error Matrix

Error matrix $E_{ij} = (H^{-1})_{ij}$

e.g.



or



N.B. ERROR NOT GIVEN BY $\ell = \ell_{max} - \frac{1}{2}$ WHEN VARYING x FROM BEST VALUE WHILE KEEPING Y.... CONSTANT

ERROR IS GIVEN BY $\ell = \ell_{max} - \frac{1}{2}$ WHEN VARYING x FROM BEST VALUE WHILE .....

PROFILE $\mathcal{L}$

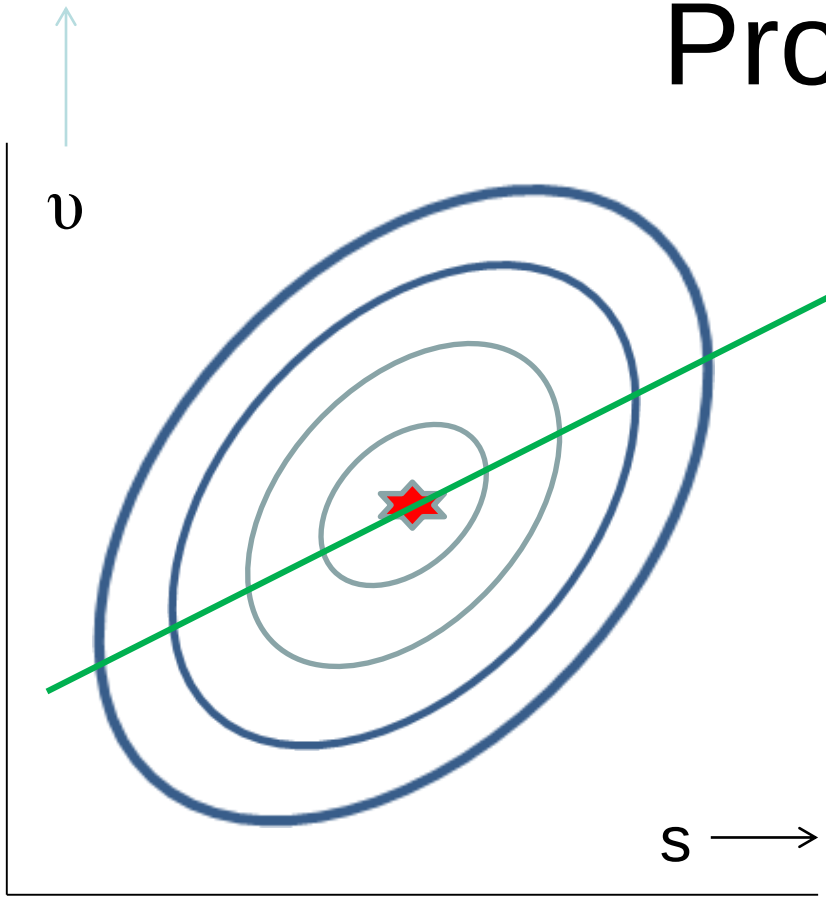
$\mathcal{L}_{prof} = \mathcal{L}(\beta, \nu_{best}(\beta))$, where

β = param of interest

ν = nuisance param(s)

Uncertainty on β from decrease in $\ln(\mathcal{L}_{prof})$ by 0.5

Profile $\mathcal{L}$

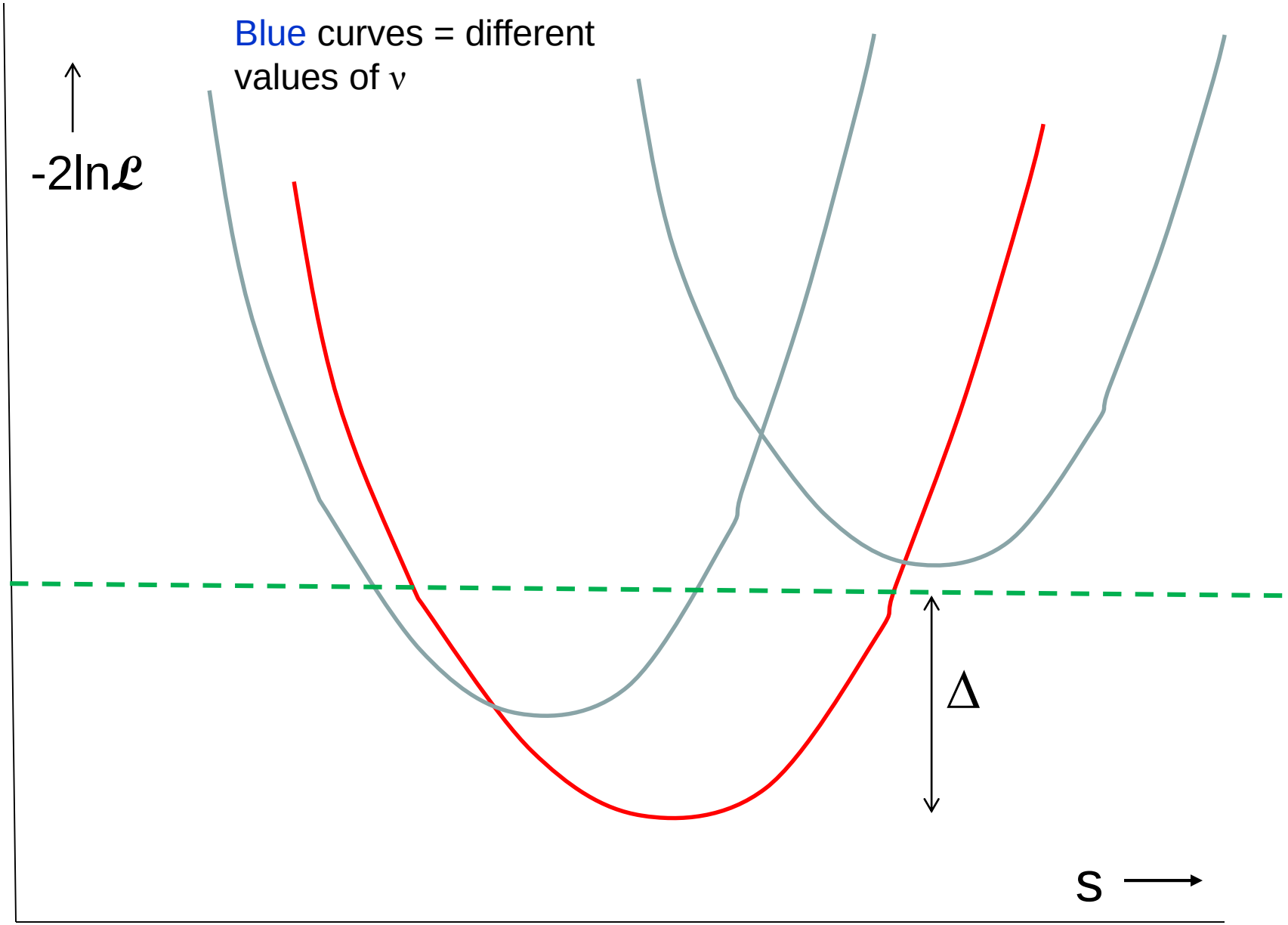


Contours of ln$\mathcal{L}$(s,υ)
s = physics param
υ = nuisance param

A method for dealing with systematics
Stat uncertainty on s from width of $\mathcal{L}$ fixed at υ_{best}

Total uncertainty on s from width of $\mathcal{L}(s,\upsilon_{prof}(s)) = \mathcal{L}_{prof}$
$\upsilon_{prof}(s)$ is best value of υ at that s
$\upsilon_{prof}(s)$ as fn of s lies on green line

Total uncert ≥ stat uncertainty


Blue curves = different values of ν
-2lnℒ
Δ
S

Extended Maximum Likelihood

Maximum Likelihood uses shape → parameters
Extended Maximum Likelihood uses shape and normalisation
i.e. EML uses prob of observing:

a) sample of N events; and
b) given data distribution in x,……
→ shape parameters and normalisation.

Example 1: Angular distribution

Observe N events total e.g 100
F forward 96
B backward 4

Rate estimates	ML	EML
Total	-----	100±10
Forward	96±2	96±10
Backward	4±2	4± 2

ML and EML

ML uses fixed (data) normalisation
EML has normalisation as parameter

Example 2: Cosmic ray experiment

	See 96 protons	and	4 heavy nuclei
ML estimate	96 ± 2% protons		4 ±2% heavy nuclei
EML estimate	96 ± 10 protons		4 ± 2 heavy nuclei

Example 3: Decay of resonance

Use ML for Branching Ratios
Use EML for Partial Decay Rates

Relation between Poisson and Binomial

N people in lecture, m males and f females (N = m + f)

Assume these are representative of basic rates: ν people νp males $\nu(1-p)$ females

Probability of observing N people = $P_{Poisson} = e^{-\nu} \nu^{N} / N!$

Prob of given male/female division = $P_{Binom} = \frac{N!}{m!\, f!} p^{m} (1-p)^{f}$

Prob of N people, m male and f female = $P_{Poisson} P_{Binom}$

$$= \frac{e^{-\nu p} \nu^{m} p^{m}}{m!} * \frac{e^{-\nu(1-p)} \nu^{f} (1-p)^{f}}{f!}$$

= Poisson prob for males * Poisson prob for females

People	Male	Female
Patients	Cured	Remain ill
Decaying nuclei	Forwards	Backwards
Cosmic rays	Protons	Other particles

	Moments	Max Like	Least squares
Easy?	Yes, if…	Normalisation, maximisation messy	Minimisation
Efficient?	Not very	Usually best	Sometimes = Max Like
Input	Separate events	Separate events	Histogram
Goodness of fit	Messy	No (unbinned)	Easy
Constraints	No	Yes	Yes
N dimensions	Easy if ….	Norm, max messier	Easy
Weighted events	Easy	Errors difficult	Easy
Bgd subtraction	Easy	Troublesome	Easy
Inverse Covariance Matrix	Observed spread, or analytic	$\left\{ -\frac{\partial^2 \ell}{\partial p_i \partial p_j} \right\}$	$\left\{ \frac{\partial^2 S}{2 \partial p_i \partial p_j} \right\}$
Main feature	Easy	Best	Goodness of Fit

DO'S AND DONT'S WITH $\mathcal{L}$

- NORMALISATION FOR LIKELIHOOD
- JUST QUOTE UPPER LIMIT
- $\Delta(\ln \mathcal{L}) = 0.5$ RULE ←
- $\mathcal{L}_{max}$ AND GOODNESS OF FIT ←
- $\int_{p_L}^{p_U} \mathcal{L}\, dp = 0.90$
- BAYESIAN SMEARING OF $\mathcal{L}$
- USE CORRECT $\mathcal{L}$ (PUNZI EFFECT)

$\Delta \ln \mathcal{L} = -1/2$ rule

If $\mathcal{L}(\mu)$ is Gaussian, following definitions of σ are equivalent:

1) RMS of $\mathcal{L}(\mu)$

2) $1/\sqrt{(-d^2 \ln \mathcal{L}/d\mu^2)}$

3) $\ln(\mathcal{L}(\mu_0 \pm \sigma) = \ln(\mathcal{L}(\mu_0)) - 1/2$

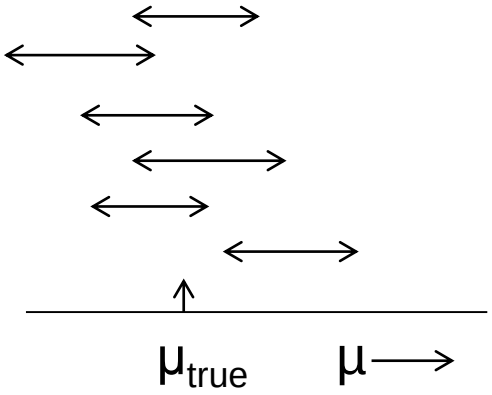
If $\mathcal{L}(\mu)$ is non-Gaussian, these are no longer the same

~~"Procedure 3) above still gives interval that contains the true value of parameter μ with 68% probability"~~

Heinrich: CDF note 6438 (see CDF Statistics Committee Web-page)

Barlow: Phystat05

COVERAGE



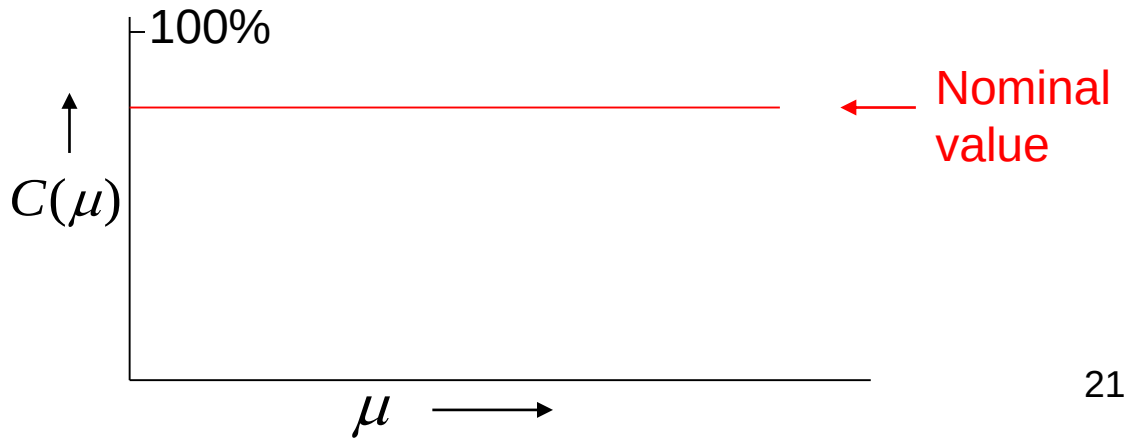
How often does quoted range for parameter include param's true value?

N.B. Coverage is a property of METHOD, not of a particular exptl result

Coverage can vary with μ

Study coverage of different methods of Poisson parameter μ, from observation of number of events n

Hope for:



COVERAGE

If true for all μ : "correct coverage"

P< α for some μ "undercoverage"
(this is serious !)

P> α for some μ "overcoverage"

Conservative

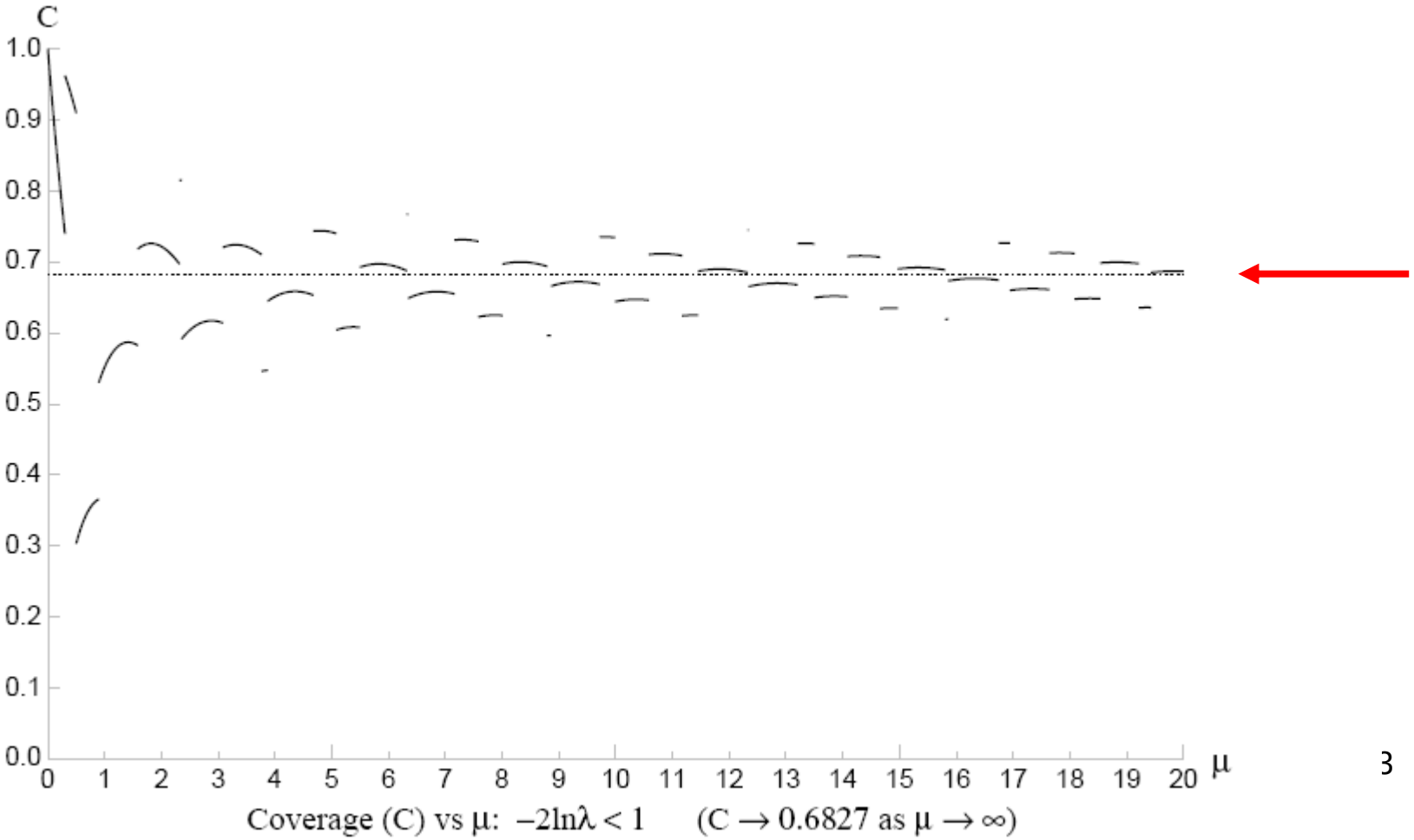
Loss of rejection power

Some Bayesians regard Coverage as irrelevant

Coverage : $\mathcal{L}$ approach (Not Neyman construction)

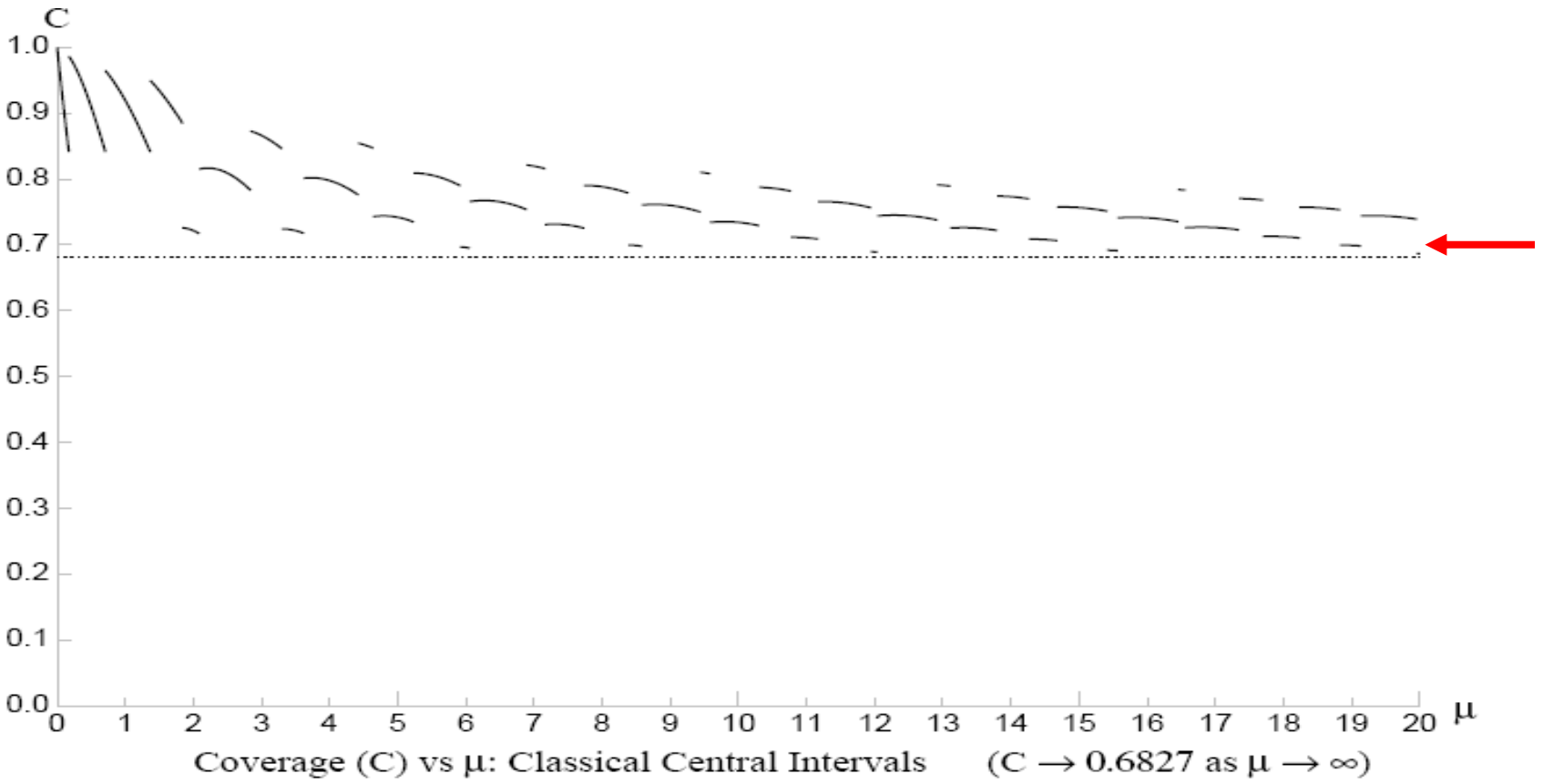
$P(n,\mu) = e^{-\mu}\mu^n/n!$ (Joel Heinrich CDF note 6438)

$-2\ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



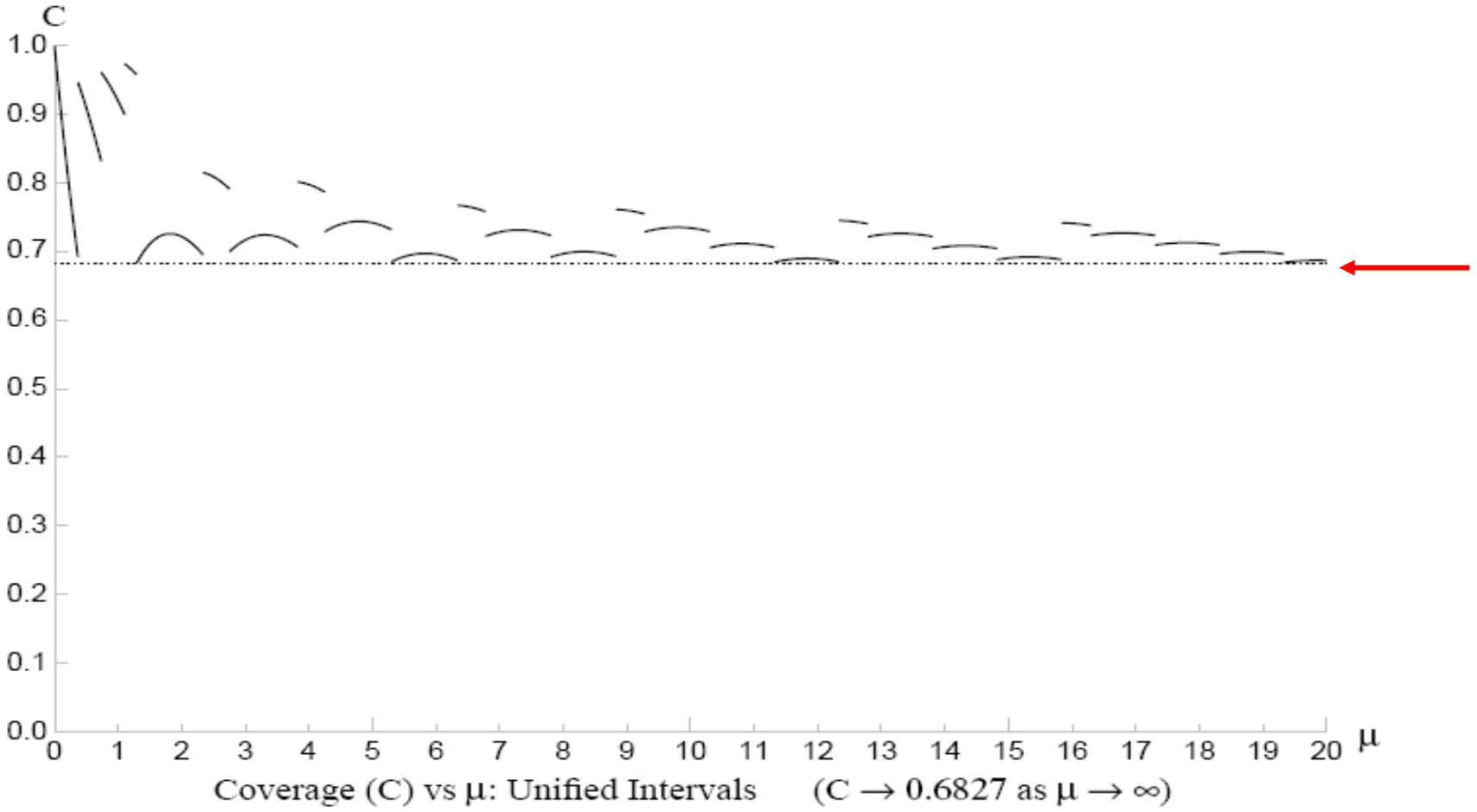
Neyman central intervals, NEVER undercover

(Conservative at both ends)



Feldman-Cousins Unified intervals

Neyman construction so NEVER undercovers



Coverage (C) vs μ: Unified Intervals ($C \to 0.6827$ as $\mu \to \infty$)

Unbinned $\mathcal{L}_{max}$ and Goodness of Fit?

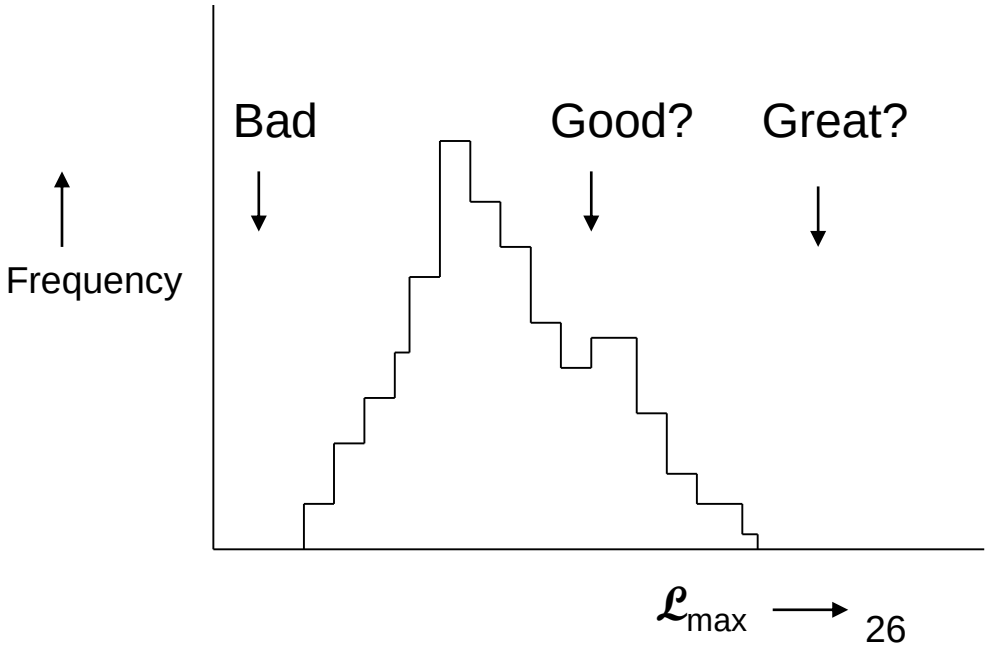
Find params by maximising $\mathcal{L}$

So larger $\mathcal{L}$ better than smaller $\mathcal{L}$

So $\mathcal{L}_{max}$ gives Goodness of Fit??

Monte Carlo distribution

of unbinned $\mathcal{L}_{max}$



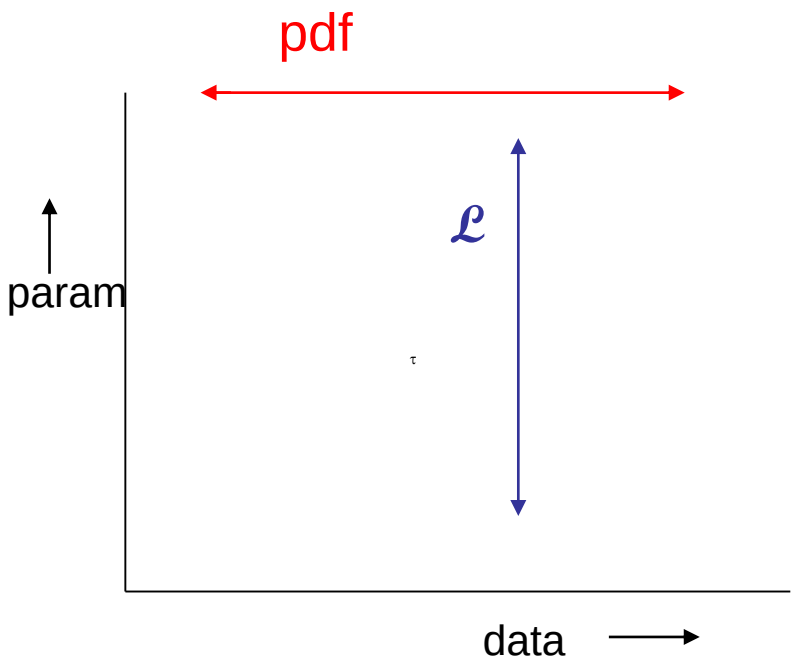
Not necessarily:

$\mathcal{L}$(data, params)

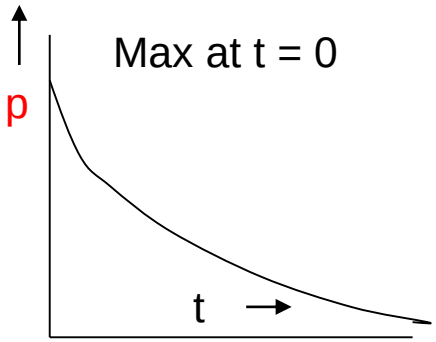
fixed vary

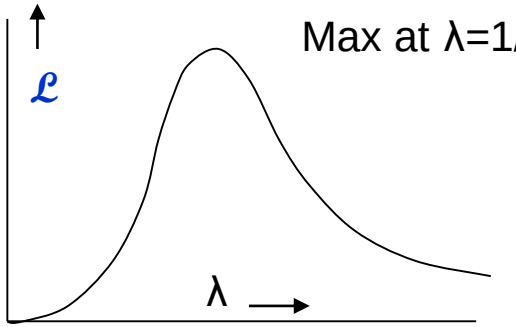
Contrast pdf(data, params)

vary fixed



e.g. $p(\lambda) = \lambda \exp(-\lambda t)$





Example 1

Fit exponential to times t_1, t_2 , t_3 ……. [Joel Heinrich, CDF 5639]

$\mathcal{L} = \prod \lambda \exp(-\lambda t_i)$

$\ln \mathcal{L}_{max} = -N(1 + \ln t_{av})$

i.e. Depends only on AVERAGE t, but is

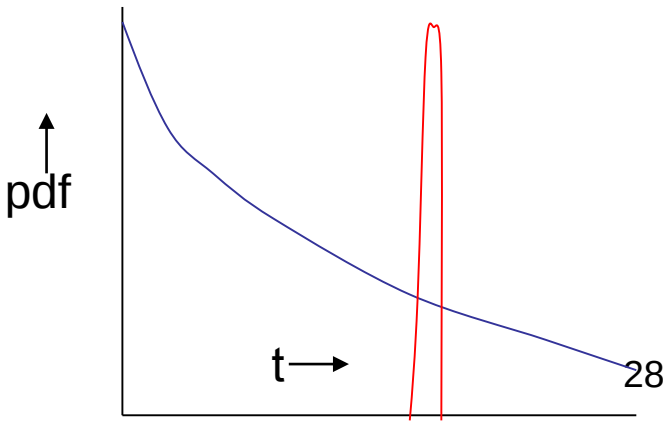
INDEPENDENT OF DISTRIBUTION OF t (except for……..)

(Average t is a sufficient statistic)

Variation of $\mathcal{L}_{max}$ in Monte Carlo is due to variations in samples' average t , but

NOT TO BETTER OR WORSE FIT

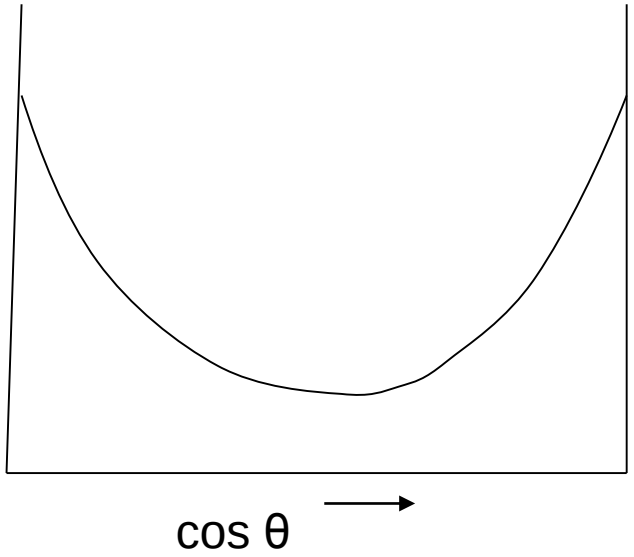
Same average t ⇒ same $\mathcal{L}_{max}$



Example 2

$$\frac{dN}{d\cos\theta} = \frac{1+\alpha\cos^2\theta}{1+\alpha/3}$$

$$\mathcal{L} = \prod_i \frac{1+\alpha\cos^2\vartheta_i}{1+\alpha/3}$$



pdf (and likelihood) depends only on $\cos^2\theta_i$

Insensitive to sign of $\cos\theta_i$

So data can be in very bad agreement with expected distribution

e.g. all data with $\cos\theta < 0$

and $\mathcal{L}_{max}$ does not know about it.

Example of general principle

$\mathcal{L}_{max}$ and Goodness of Fit?

Conclusion:

$\mathcal{L}$ has sensible properties with respect to parameters

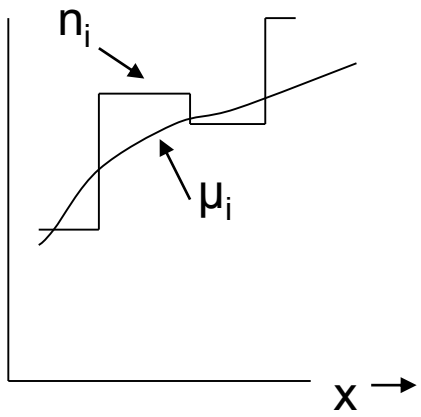
NOT with respect to data

$\mathcal{L}_{max}$ within Monte Carlo peak is NECESSARY

not SUFFICIENT

('Necessary' doesn't mean that you have to do it!)

Binned data and Goodness of Fit using $\mathcal{L}$-ratio



$$\mathcal{L} = \prod_i p_{ni}(\mu_i)$$

$$\mathcal{L}_{best} = \prod_i p_{ni}(\mu_{i,best}) = \prod_i p_{ni}(n_i)$$

$\ln[\mathcal{L}\text{-ratio}] = \ln[\mathcal{L}/\mathcal{L}_{best}]$

$\xrightarrow{\text{large } \mu_i}$ $-0.5\chi^2$ i.e. Goodness of Fit

$\mathcal{L}_{best}$ is independent of parameters of fit,

and so same parameter values from $\mathcal{L}$ or $\mathcal{L}$-ratio

Baker and Cousins, NIM A221 (1984) 437

Conclusions

How it works, and how to estimate uncertainties

Likelihood or Extended Likelihood

Several Parameters

$\mathcal{L}$ikelihood does not guarantee coverage

Unbinned $\mathcal{L}_{max}$ and Goodness of Fit

Getting $\mathcal{L}$ wrong: Punzi effect

Giovanni Punzi @ PHYSTAT2003

"Comments on $\mathcal{L}$ fits with variable resolution"

Separate two close signals, when resolution σ varies event by event, and is different for 2 signals

e.g. 1) Signal 1 $1+\cos^2\theta$

Signal 2 Isotropic

and different parts of detector give different σ

2) M (or τ)

Different numbers of tracks → different σ_M (or σ_τ)

Events characterised by x_i and σ_i

A events centred on x = 0

B events centred on x = 1

$\mathcal{L}(f)_{wrong} = \Pi\,[f * G(x_i,0,\sigma_i) + (1-f) * G(x_i,1,\sigma_i)]$

$\mathcal{L}(f)_{right} = \Pi\,[f*p(x_i,\sigma_i;A) + (1-f) * p(x_i,\sigma_i;B)]$

$$p(S,T) = p(S|T) * p(T)$$

$$p(x_i,\sigma_i|A) = p(x_i|\sigma_i,A) * p(\sigma_i|A)$$

$$= G(x_i,0,\sigma_i) * p(\sigma_i|A)$$

So

$\mathcal{L}(f)_{right} = \Pi[f * G(x_i,0,\sigma_i) * p(\sigma_i|A) + (1-f) * G(x_i,1,\sigma_i) * p(\sigma_i|B)]$

If $p(\sigma|A) = p(\sigma|B)$, $\mathcal{L}_{right} = \mathcal{L}_{wrong}$

but NOT otherwise

Punzi's Monte Carlo for A : $G(x,0,\sigma_A)$

B : $G(x,1,\sigma_B)$

$f_A = 1/3$

		$\mathcal{L}_{wrong}$		$\mathcal{L}_{right}$	
σ_A	σ_B	f_A	σ_f	f_A	σ_f
1·0	1·0	0·336(3)	0·08	Same	
1·0	1·1	0·374(4)	0·08	0· 333(0)	0
1·0	2·0	0·645(6)	0·12	0·333(0)	0
1 → 2	1.5 →3	0·514(7)	0·14	0·335(2)	0·03
1.0	1 → 2	0.482(9)	0.09	0.333(0)	0

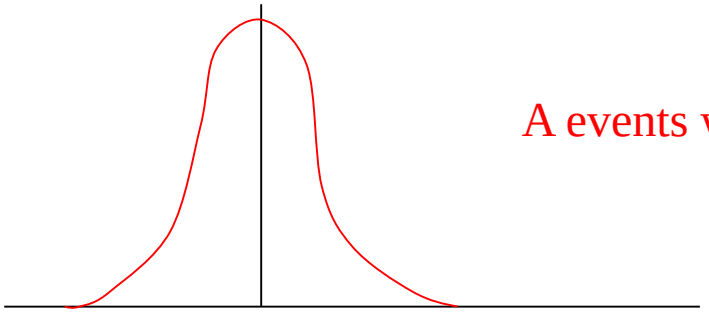
1) $\mathcal{L}_{wrong}$ OK for $p(\sigma_A) = p(\sigma_B)$, but otherwise BIASSED

2) $\mathcal{L}_{right}$ unbiassed, but $\mathcal{L}_{wrong}$ biassed (enormously)!

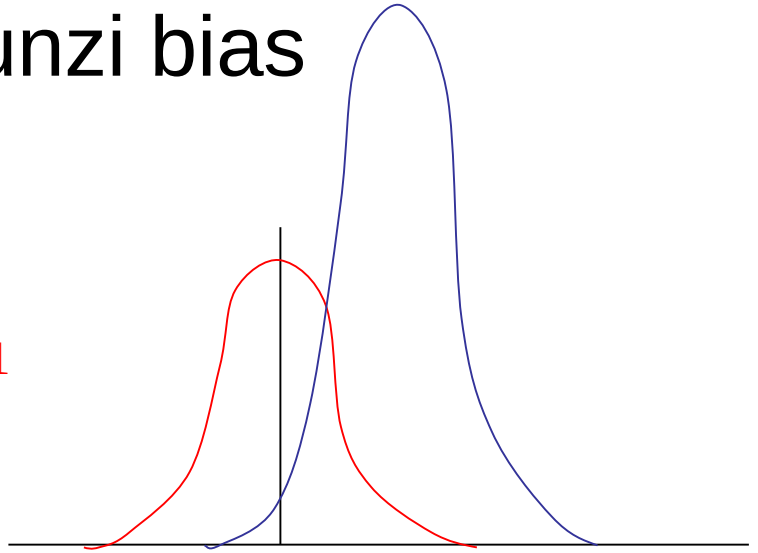
3) $\mathcal{L}_{right}$ gives smaller σ_f than $\mathcal{L}_{wrong}$

Explanation of Punzi bias

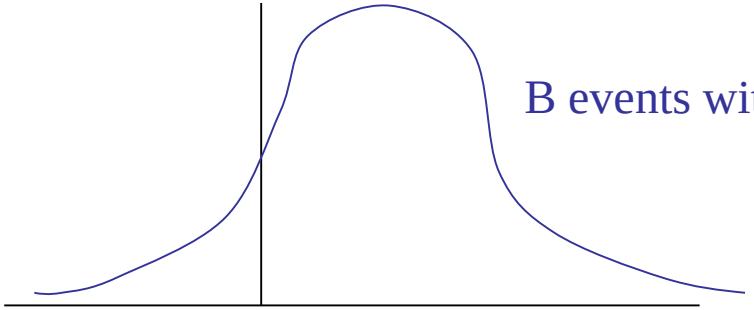
$\sigma_A = 1$ $\sigma_B = 2$



A events with $\sigma = 1$

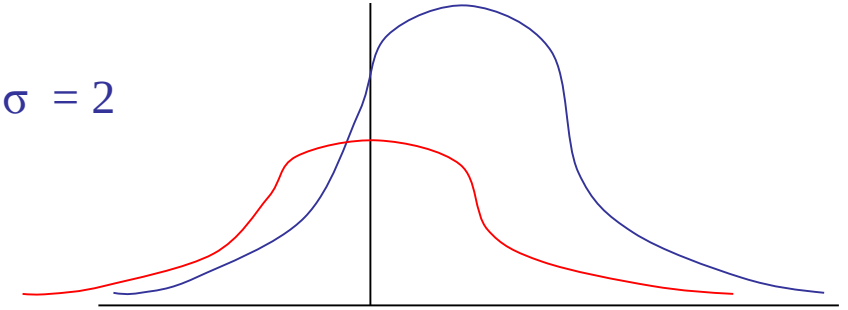


B events with $\sigma = 2$



x →

ACTUAL DISTRIBUTION



x →

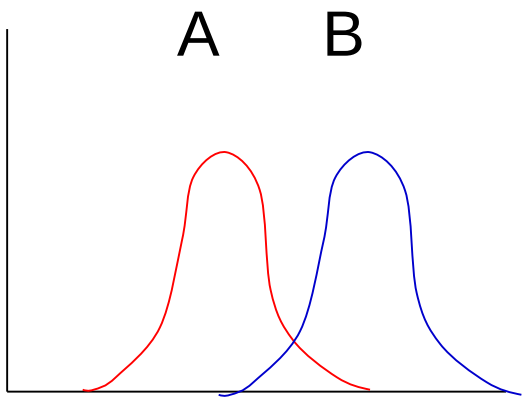
FITTING FUNCTION

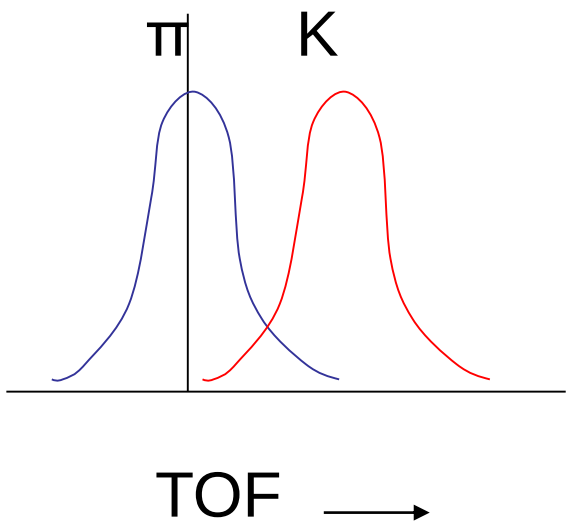
[N_A/N_B variable, but same for A and B events]

Fit gives upward bias for N_A/N_B because (i) that is much better for A events; and

(ii) it does not hurt too much for B events

Another scenario for Punzi problem: PID





Originally:

Positions of peaks = constant

K-peak → π-peak at large momentum

σ_i variable, $(\sigma_i)_A \neq (\sigma_i)_B$

$\sigma_i \sim$ constant, $p_K \neq p_\pi$

COMMON FEATURE: Separation/Error ≠ Constant

Where else??

MORAL: Beware of event-by-event variables whose pdf's do not appear in $\mathcal{L}$

Avoiding Punzi Bias

BASIC RULE:
Write pdf for ALL observables, in terms of parameters

- Include $p(\sigma|A)$ and $p(\sigma|B)$ in fit
  (But then, for example, particle identification may be determined more by momentum distribution than by PID)

OR

- Fit each range of σ_i separately, and add $(N_A)_i \rightarrow (N_A)_{total}$, and similarly for B

Incorrect method using $\mathcal{L}_{wrong}$ uses weighted average of $(f_A)_j$, assumed to be independent of j

Talk by Catastini at PHYSTAT05