$\mathcal{L}$ikelihoods

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

Louis Lyons

Oxford & Imperial College

CMS

Aachen

March 2023

Topics

What it is

How it works: Resonance

Uncertainty estimates

Coverage

Several Parameters

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

COMBINING PROFILE LIKELIHOODS
NORMALISATION FOR LIKELIHOOD
$\Delta(\ln \mathcal{L}) = 0.5$ RULE
$\mathcal{L}_{max}$ AND GOODNESS OF FIT

Bayes and Frequentism: What is Probability?

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
- BAYESIAN SMEARING OF $\mathcal{L}$
- USE CORRECT $\mathcal{L}$ (PUNZI EFFECT)

Simple example: Parameter for Angular distribution

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

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

= 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

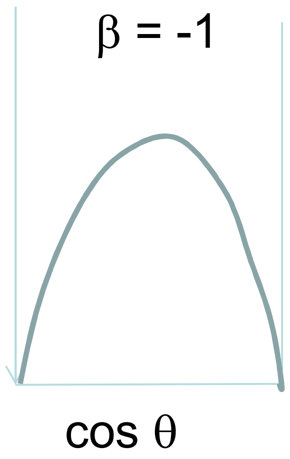
Best estimate of β is that which maximises $\mathcal{L}$

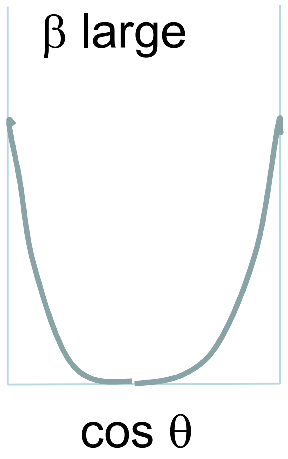
Values of β for which $\mathcal{L}$ is very small are ruled out

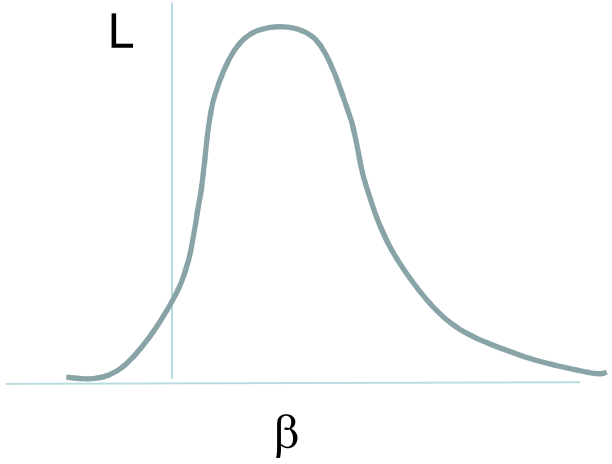
Precision of estimate for β comes from width of $\mathcal{L}$ distribution

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

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

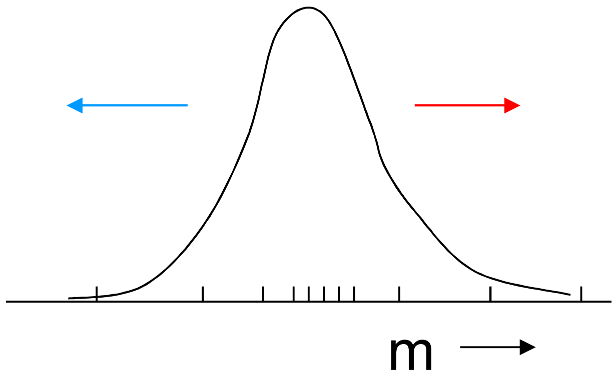




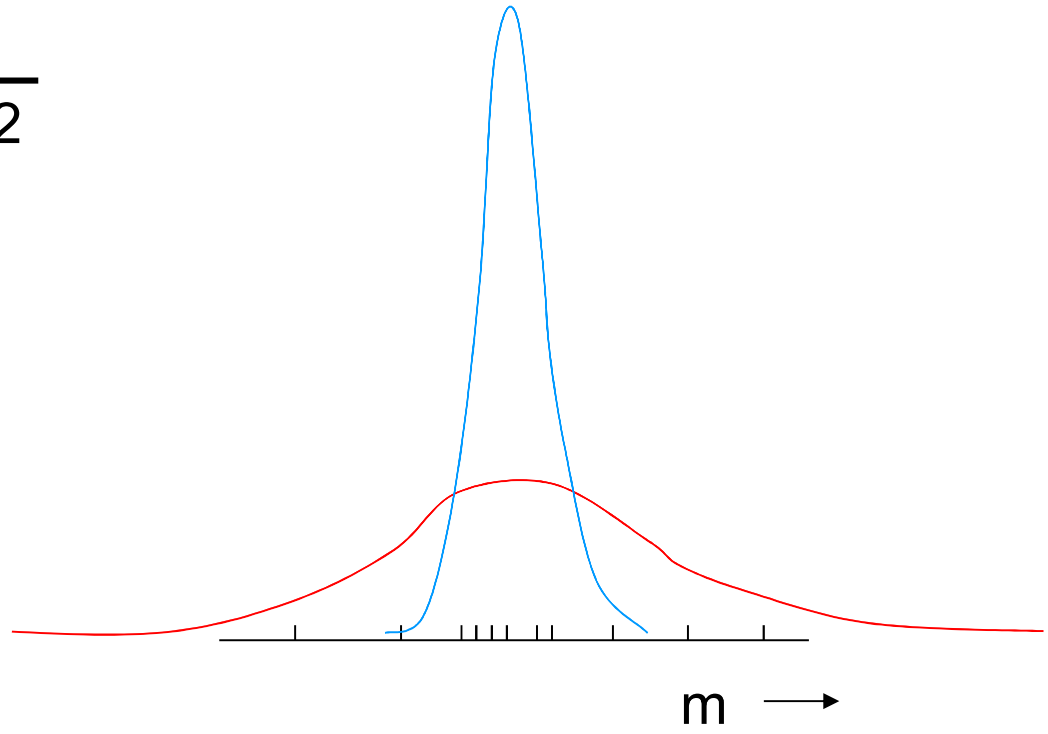


How it works: Resonance

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



Vary M_0



Vary Γ

Find overall optimum by allowing both to vary simultaneously

Conventional to consider

$$\ell = \ln(\mathcal{L}) = \Sigma \ln(y_i)$$

For large N, $\mathcal{L}$ → Gaussian

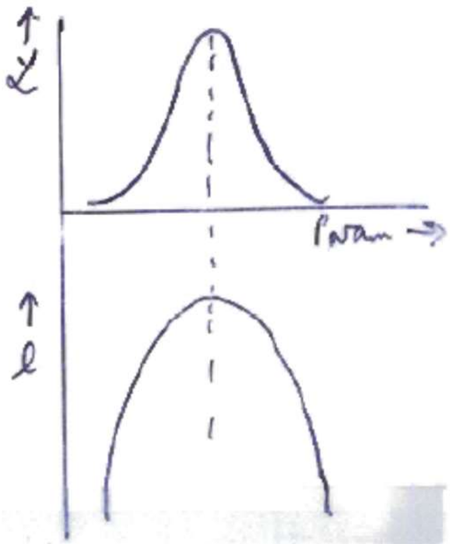
"Proof"

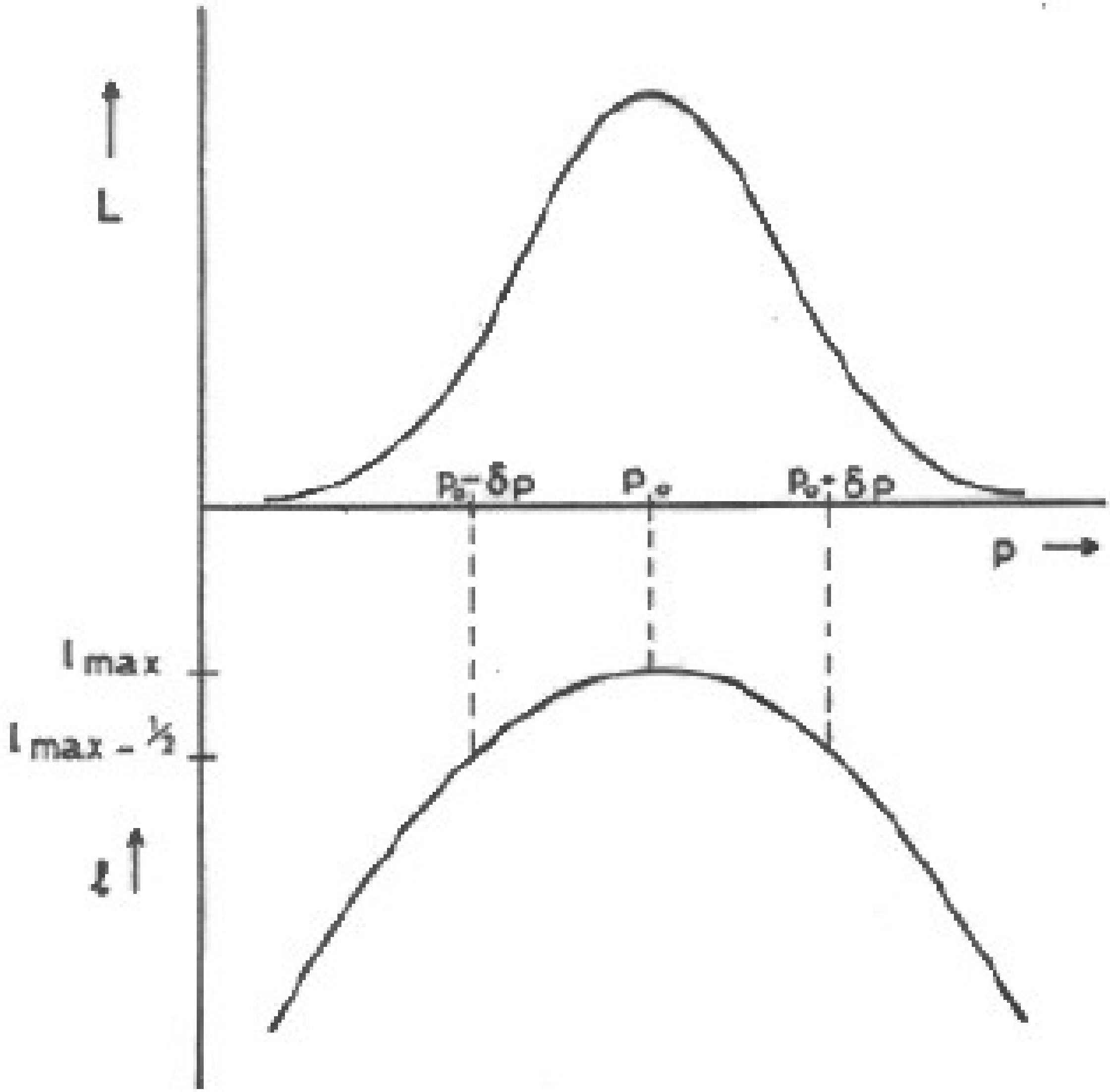
Taylor expand ℓ about its maximum

$$\ell = \ell_{max} + \frac{1}{2!}\ell''[\delta(b/a)]^2 + \ldots$$

$$= \ell_{max} - \frac{1}{2c}\delta^2 + \ldots \qquad c = -1/\ell''$$

$$\Rightarrow \mathcal{L} \sim \exp\left(-\frac{\delta^2}{2c}\right)$$




L
$p_0 - \delta p$
p_o
$p_0 + \delta p$
p
l max
l max − ½
l

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"~~

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

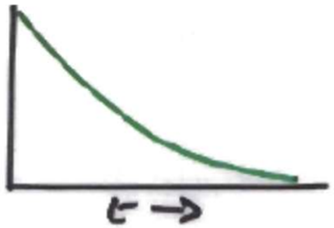
e.g 1/p rather than p; τ or λ

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

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

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

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

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

"Obvious"

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

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

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

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

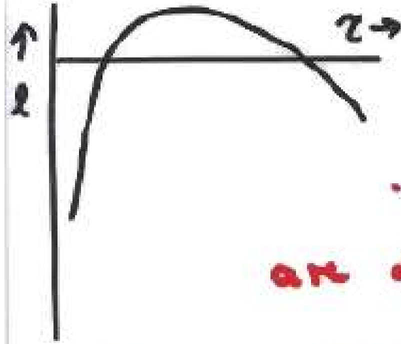
BEWARE FOR AVERAGING RESULTS

$\ln \tau - \ln \tau_{max}$ = 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 <u>not</u> on distribution of t_i

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

SEVERAL PARAMETERS

1 param p $\quad l = \ln\mathcal{L}$

. p from $dl/dp = 0$ i.e from max of $\mathcal{L}$

$$\sigma_p^2 = 1/(-d^2l/dp^2)$$

Many dimensions $l(p_1, p_2,\ldots.)$

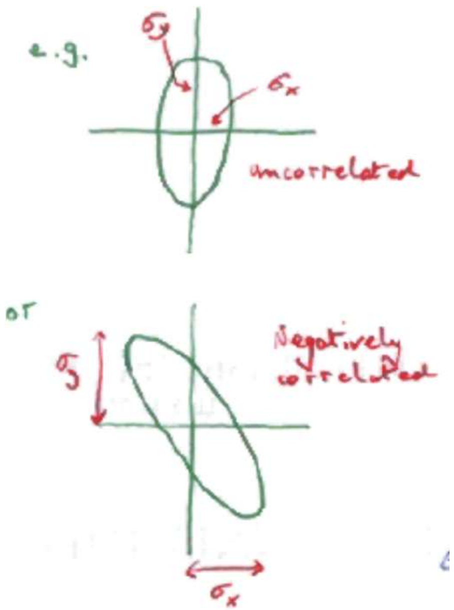
p_1, p_2, from $dl/dp_i = 0$ i.e. from max of $\mathcal{L}$

For uncertainties, define

$H_{ij} = d^2l/dp_i\,dp_j$ = Inverse Covariance Matrix

Covariance Matrix $E_{ij} = (H^{-1})_{ij}$

Diagonal Elements for variances off-diag for covariances



For many params:

N.B1 Ellipsoid with $l=l_{max}$ -0.5 does not have 68% assymptotic coverage.

N.B2 Uncert on x is not given by varying x till $l = l_{max}$ -0.5, while keeping all other params constant

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

ML and EML

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

Example 1: 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 2: Decay of resonance

Use ML for Branching Ratios
Use EML for Partial Decay Rates

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: 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

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

- COMBINING PROFILE $\mathcal{L}$s
- NORMALISATION FOR $\mathcal{L}$IKELIHOOD
- $\Delta(\ln \mathcal{L}) = 0.5$ RULE
- $\mathcal{L}_{max}$ AND GOODNESS OF FIT
- PDFs and $\mathcal{L}$IKELIHOODS

a) Max Like

Prob for fixed N = Binomial

$$= f^F (1-f)^B \frac{N!}{F!\,B!} \quad \star$$

(Prob of forwards → f)

Maximise $\ln P_a$ wrt f $\Rightarrow$ $\hat{f} = F/N$

Error on $\hat{f}$: $1/\sigma^2 = -\frac{\partial^2 \ln P_a}{\partial f^2}$

$$= \frac{N}{\hat{f}(1-\hat{f})} \qquad f = \hat{f}$$

$\Rightarrow$ Estimate of $\hat{F} = Nf = F \pm \sqrt{FB/N}$ ← Completely

- - - - - - $\hat{B} = N(1-f) = B \pm \sqrt{FB/N}$ ← anti-corr

b) EML

$$P_b = P_a \times \frac{e^{-\nu} \nu^N}{N!}$$ Poisson for overall rate

(ν = expected overall rate)

Maximise $\ln P_b(\nu, f)$

$\Rightarrow$ $\hat{\nu} = N \pm \sqrt{N}$

$\hat{f} = F/N \pm \sqrt{\frac{f(1-f)}{N}}$ } uncorrelated

For $\hat{F}$ & $\hat{B}$, either propagate errors for $\hat{F} = \hat{\nu}\hat{f}$ & $\hat{B} = \hat{\nu}(1-\hat{f})$

or rewrite eqn $\star$ as product of 2 indep Poissons

$$\left.\begin{aligned} \hat{F} &= F \pm \sqrt{F} \\ \hat{B} &= B \pm \sqrt{B} \end{aligned}\right\}$$

Danger of combining profile $\mathcal{L}$s

Experiments quote $\mathcal{L}$ikelihood, profiled over nuisance parameters, so that combinations can be performed.

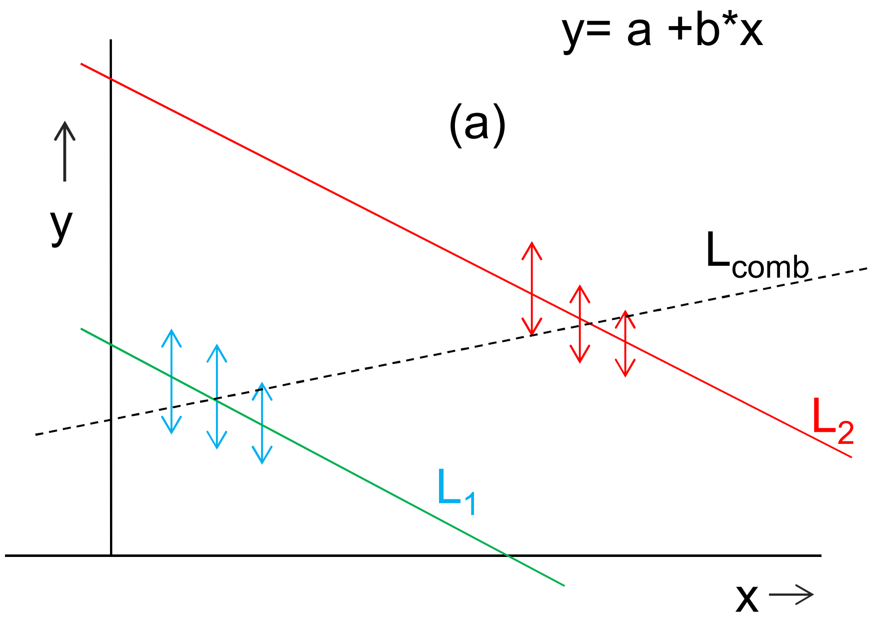
Very simple ‘tracking’ example:

* No magnetic field

* 2-D fit of straight line y = a + bx

a = parameter of interest, b = nuisance param

* Track hits in 2 subdetectors, each of 3 planes



(a) Hits in 2 sub-detectors, each with 3 planes

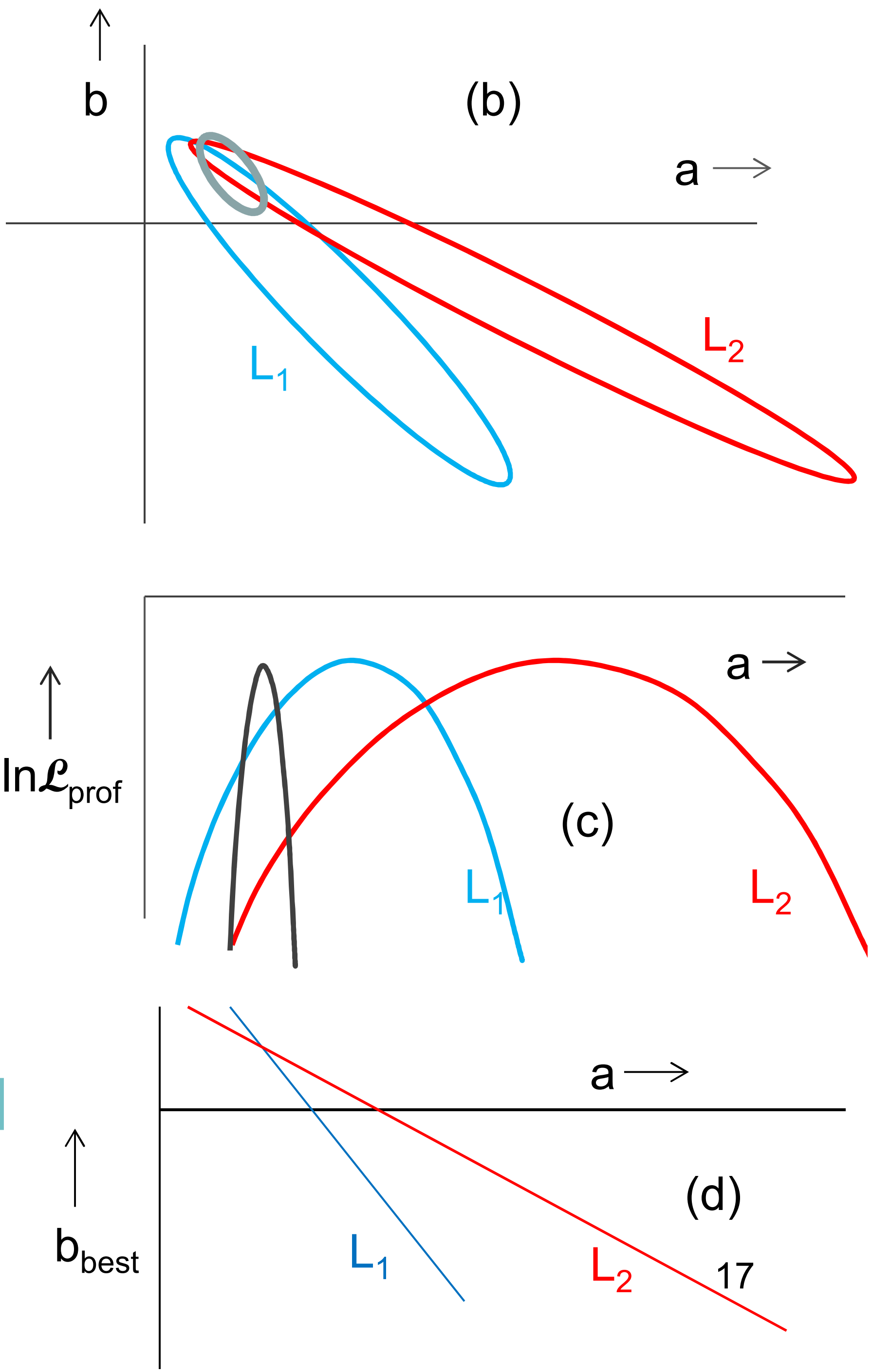
(b) Covariance ellipses for separate fits L_1 and L_2, and combined L_{comb}

(c) $\ln\mathcal{L}_{prof}$ as function of a, for all 3 lines

(d) b_{best} as a function of a

N.B. b_{best} for L_1 and L_2 are the same

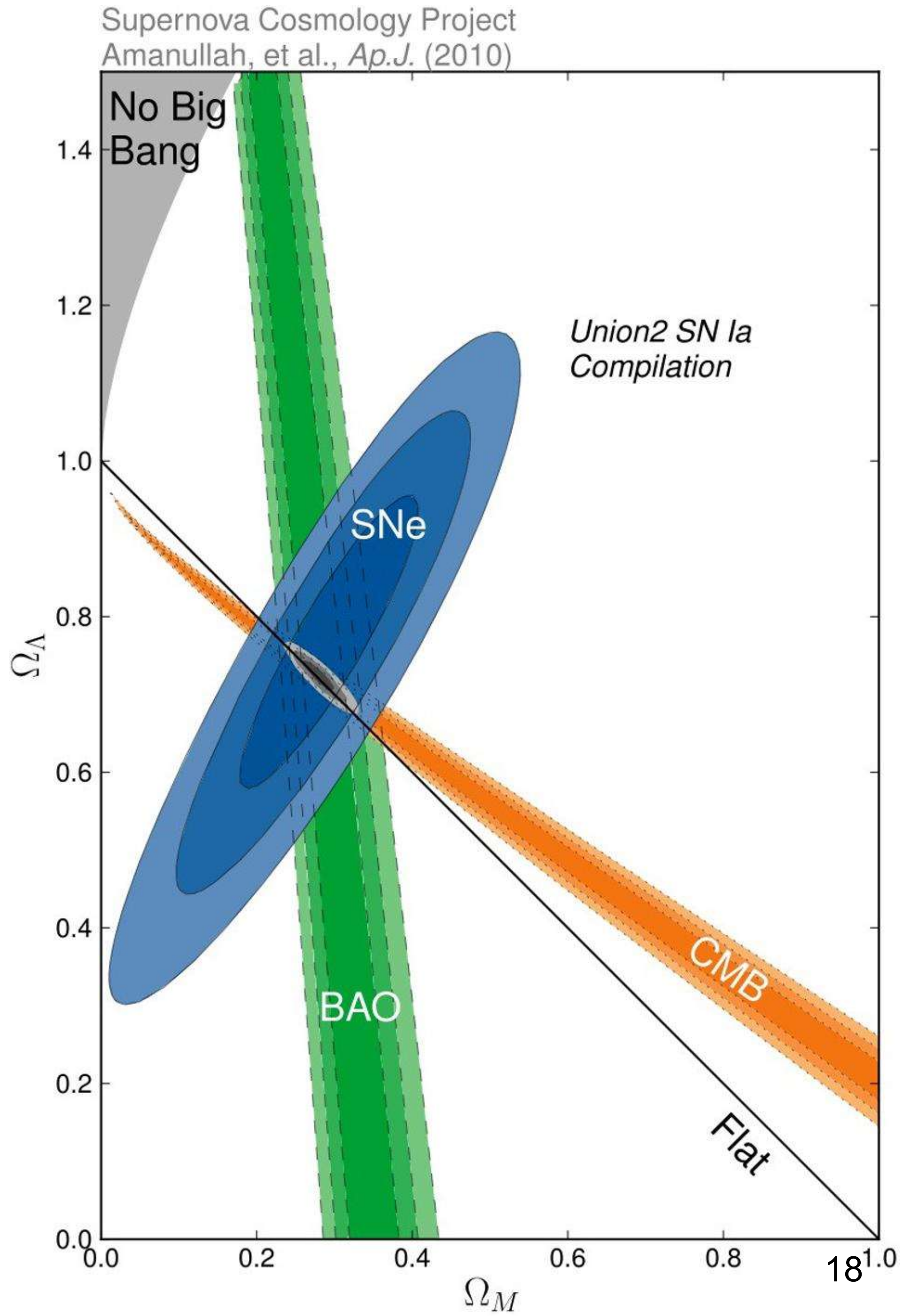
*** Combining $\mathcal{L}_{prof}$ for L_1 and L_2 loses a lot of information, and a_{best} wrong *****



COSMOLOGY EXAMPLE

Plot of dark energy fraction v dark matter fraction by various methods. Each determines dark energy fraction poorly, but combination is fine, because of different correlations.

Combining Profile Likelihoods would give very large uncertainty on dark energy fraction.



NORMALISATION FOR LIKELIHOOD

$\int P(x \mid \mu)\, dx$ MUST be independent of μ

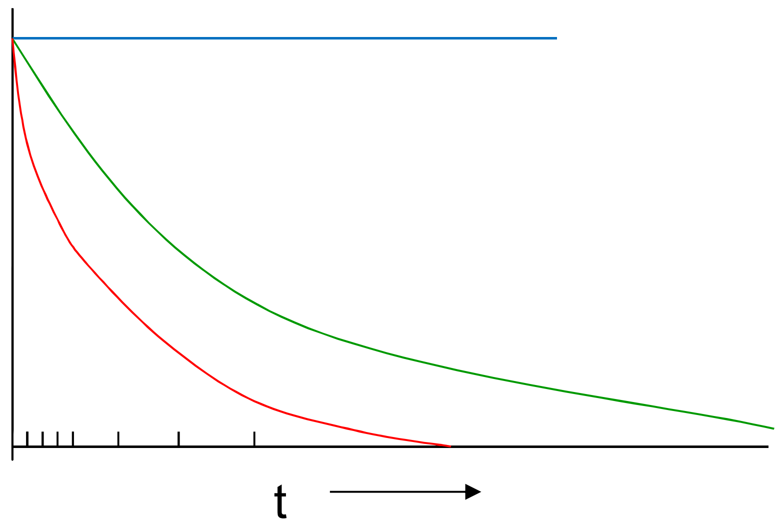
Data Param

$[\tau = \sum t_i / N]$

Exponential Distribution

INCORRECT $P(t \mid \tau) = e^{-t/\tau}$

Missing $1/\tau$



τ infinite

τ too large

τ about right

QUOTING UPPER LIMIT

"We observed no significant signal, and our 90% conf upper limit is ….."

Need to specify method e.g.

$\mathcal{L}$

Chi-squared (data or theory error)

Frequentist (Central or upper limit)

Feldman-Cousins

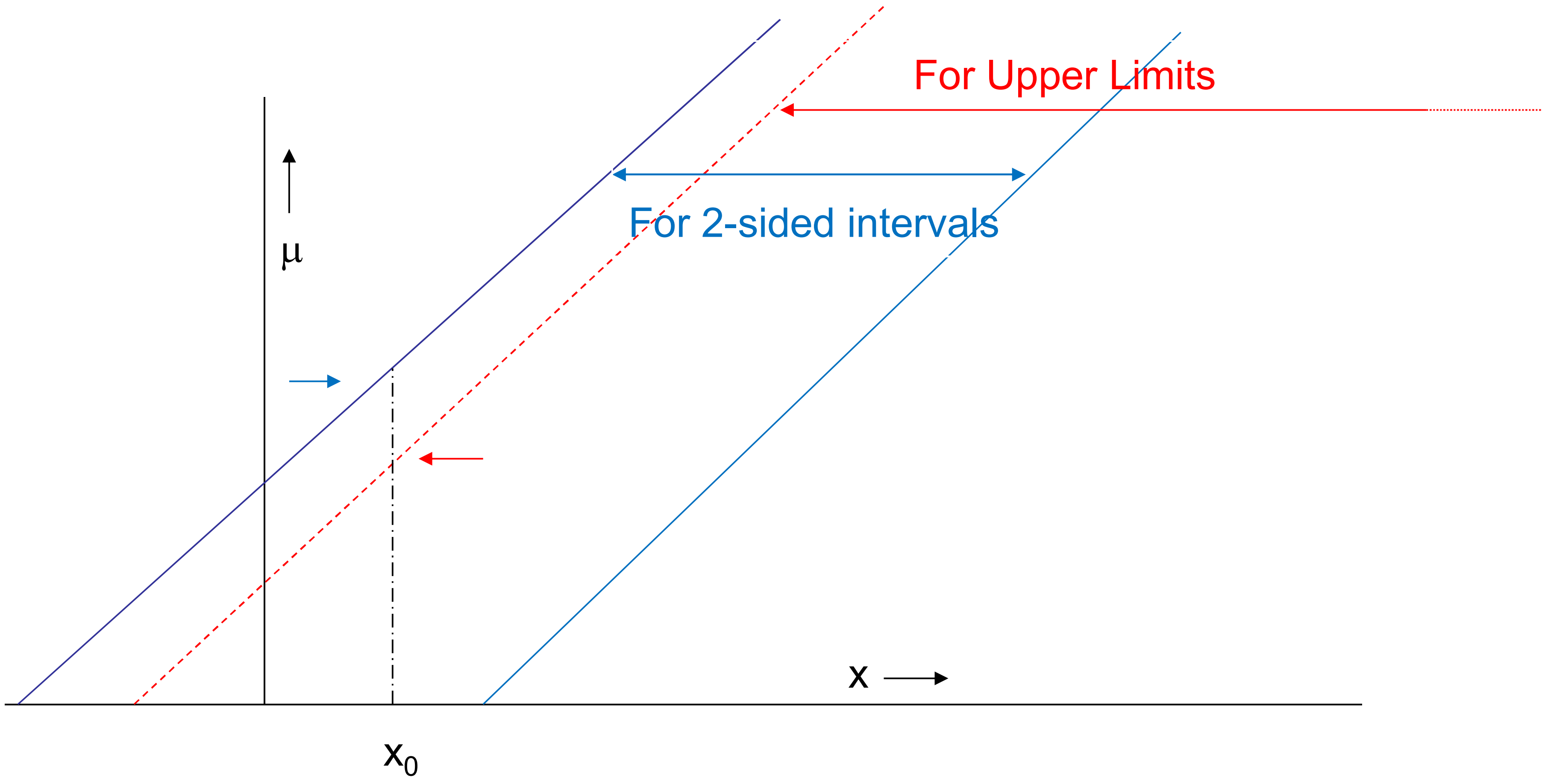
Bayes with prior = const,

"Show your $\mathcal{L}$"

1) Not always practical

2) Not sufficient for frequentist methods

90% C.L. Upper Limits



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

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

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

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

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

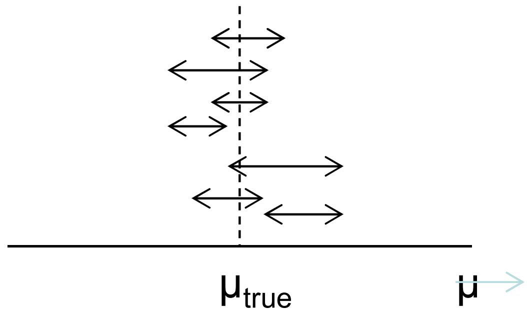
If $\mathcal{L}(\mu)$ is non-Gaussian, these are no longer the same

~~"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



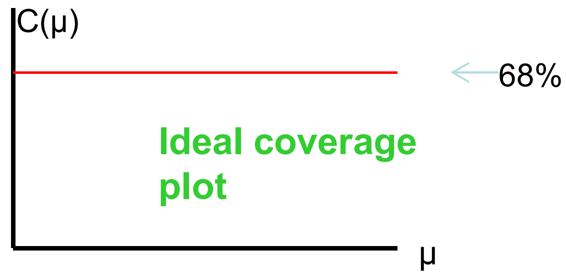
* What it is:

For given statistical method applied to many sets of data to extract confidence intervals for param μ, coverage C is fraction of ranges that contain true value of param. Can vary with μ

* Does not apply to **your** data:

It is a property of the **statistical method** used

It is **NOT** a probability statement about whether μ_{true} lies in your confidence range for μ



* Coverage plot for Poisson counting expt

Observe n counts

Estimate μ_{best} from maximum of likelihood

$\mathcal{L}(\mu) = e^{-\mu} \mu^n/n!$ and range of μ from $\ln\{\mathcal{L}(\mu_{best})/\mathcal{L}(\mu)\} < 0.5$

For each μ_{true} calculate coverage $C(\mu_{true})$, and compare with nominal 68%

Coverage

Fraction of intervals containing true value

Property of method, not of result

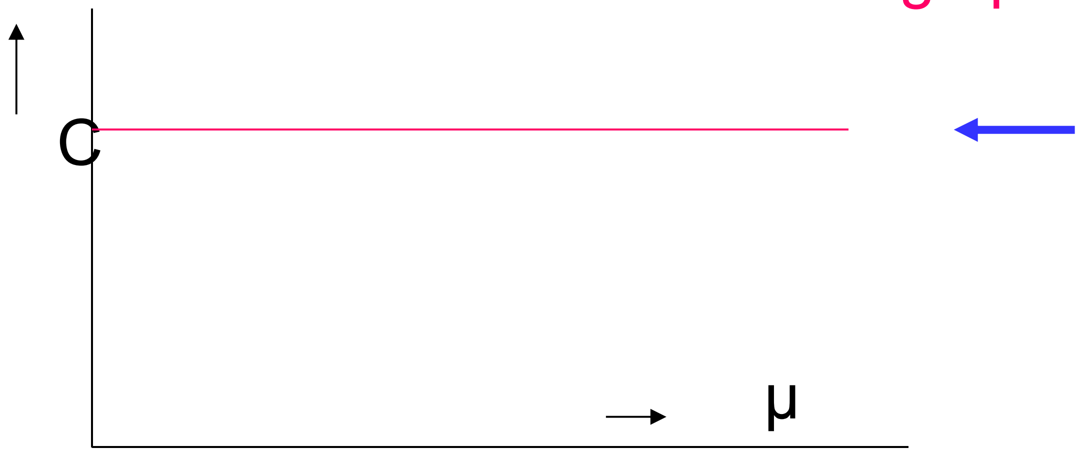
Can vary with param

Frequentist concept. Built in to Neyman construction

Some Bayesians reject idea. Coverage not guaranteed

Integer data (Poisson) → discontinuities

Ideal coverage plot



COVERAGE

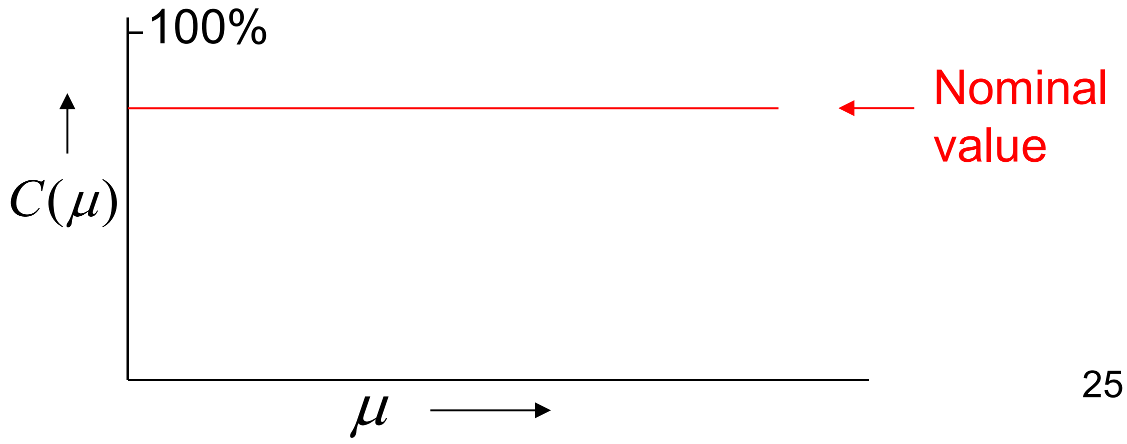
How often does quoted range for parameter include param's true value?

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"

$P < \alpha$ for some μ "undercoverage" (this is serious !)

$P > \alpha$ for some μ "overcoverage"

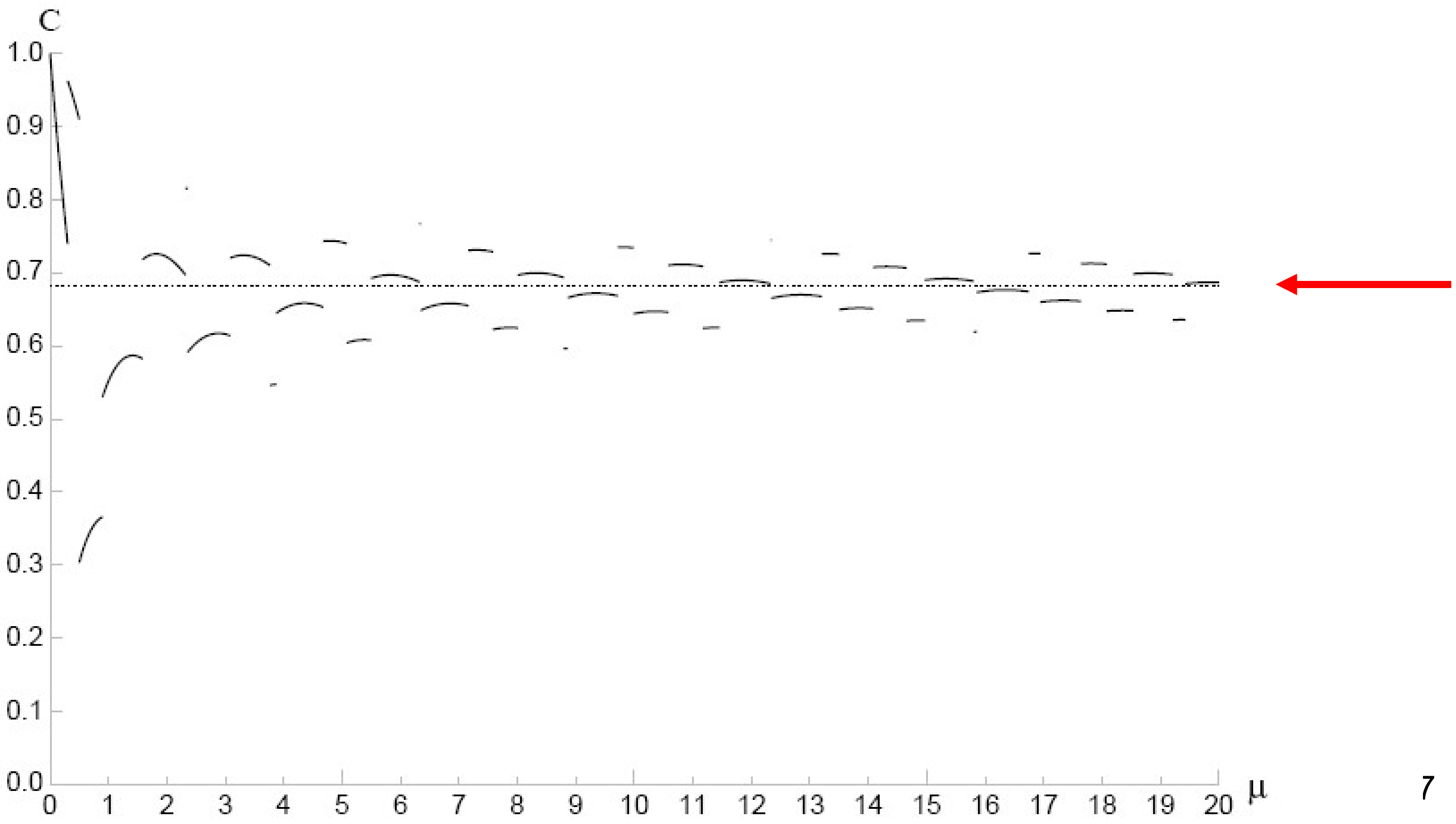
Conservative

Loss of rejection power

Coverage : $\mathcal{L}$ approach (Neyman construction)

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

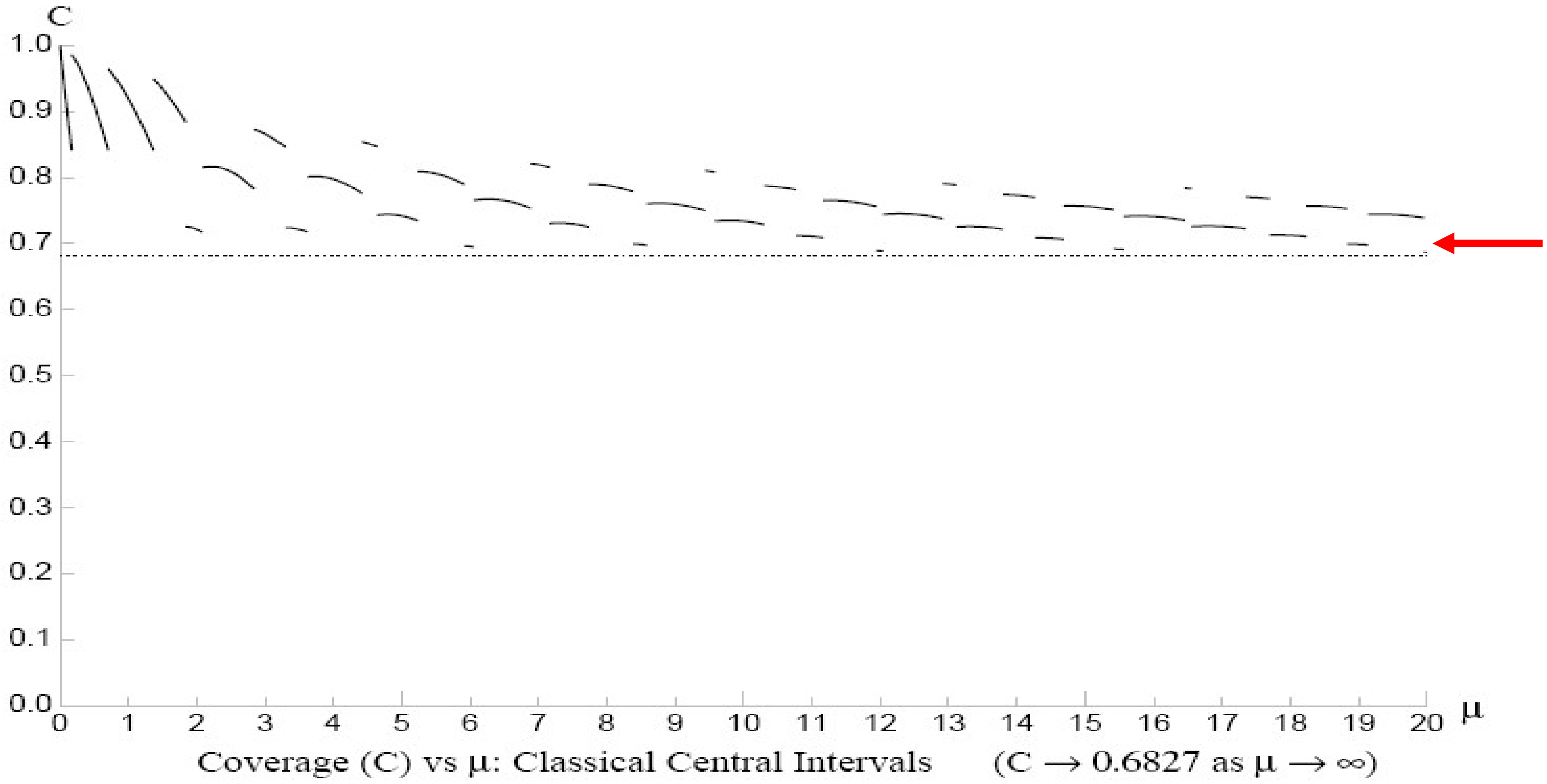
$-2 \ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



Coverage (C) vs μ: $-2\ln\lambda < 1$ ($C \rightarrow 0.6827$ as $\mu \rightarrow \infty$)

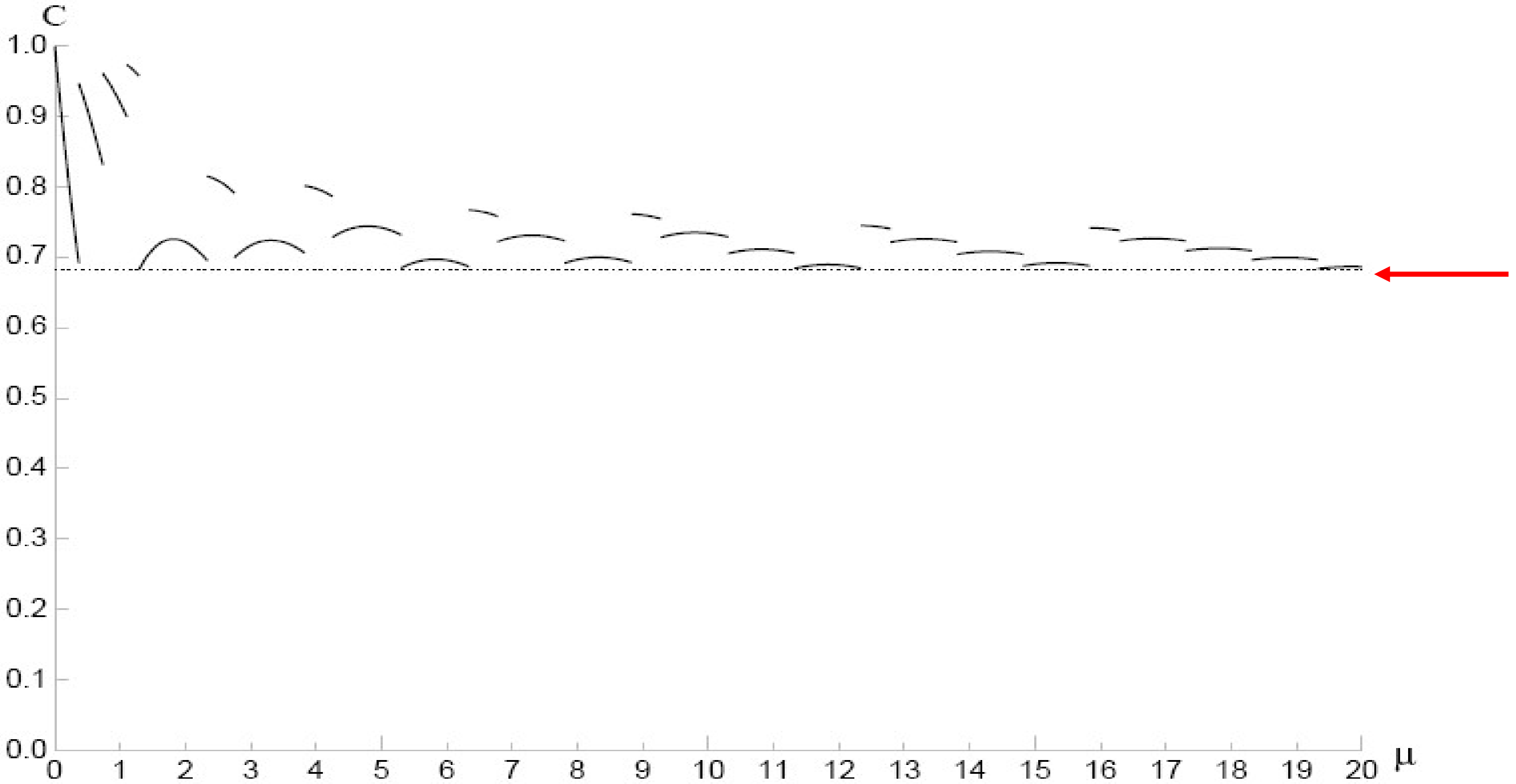
Neyman central intervals, NEVER undercover

(Conservative at both ends)



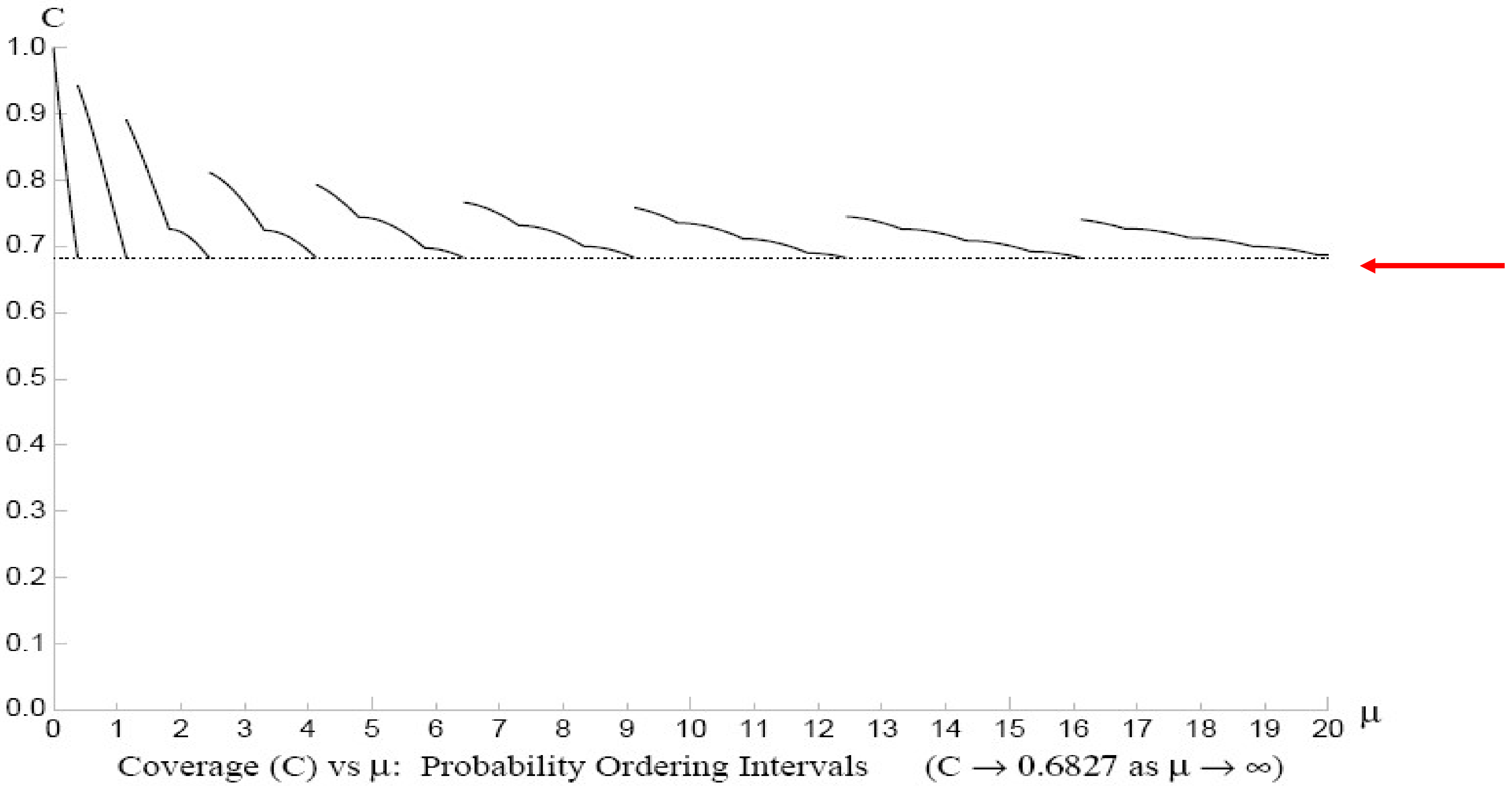
Feldman-Cousins Unified intervals

Neyman construction so NEVER undercovers



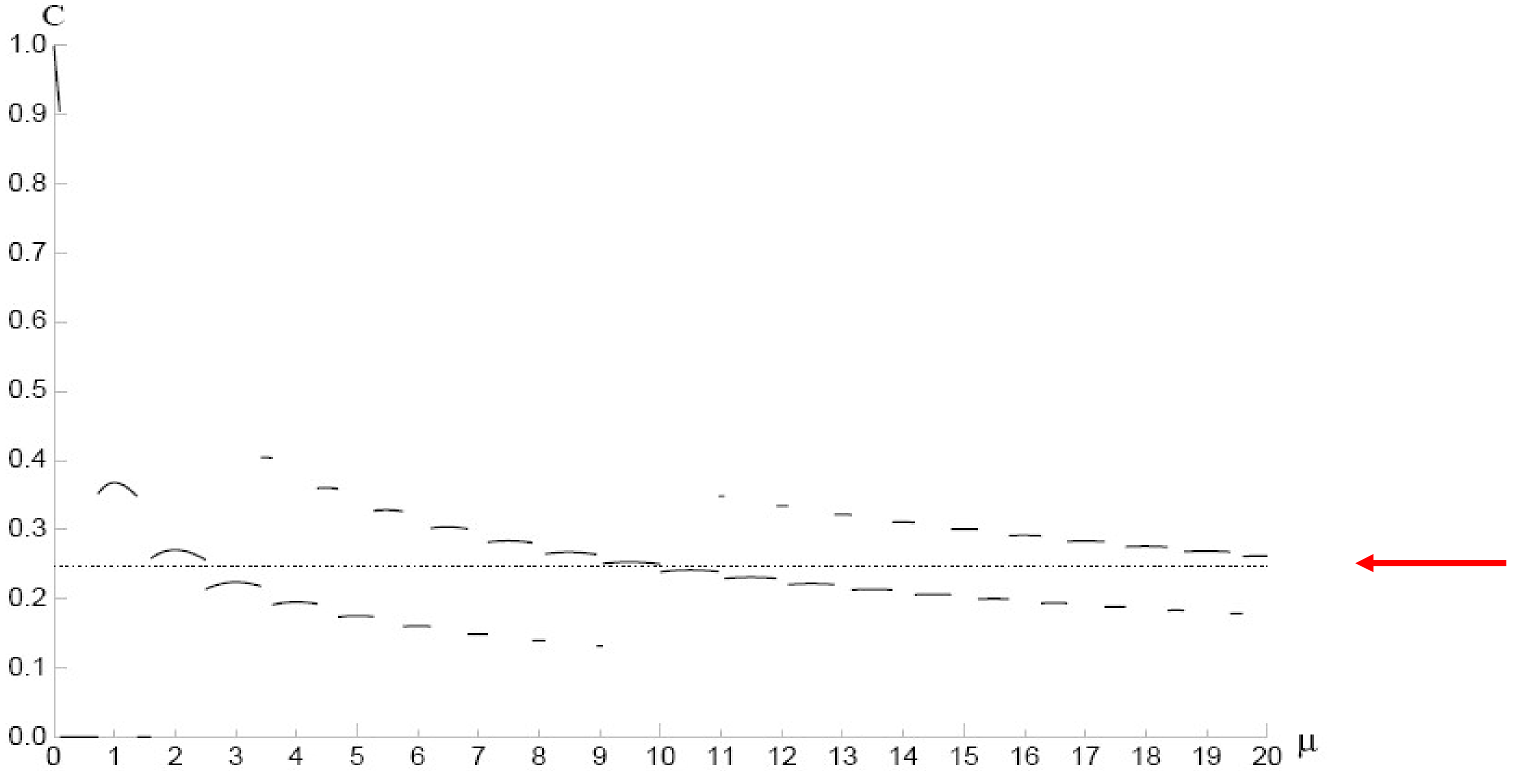
Coverage (C) vs μ: Unified Intervals (C → 0.6827 as μ → ∞)

Probability ordering



$\chi^2 = (n-\mu)^2/\mu$ $\Delta\chi^2 = 0.1$ ⟶ 24.8% coverage?

NOT Neyman : Coverage = 0% → 100%



Coverage (C) vs μ: $\chi^2 < 0.1$ ($C \rightarrow 0.2482$ as $\mu \rightarrow \infty$)

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

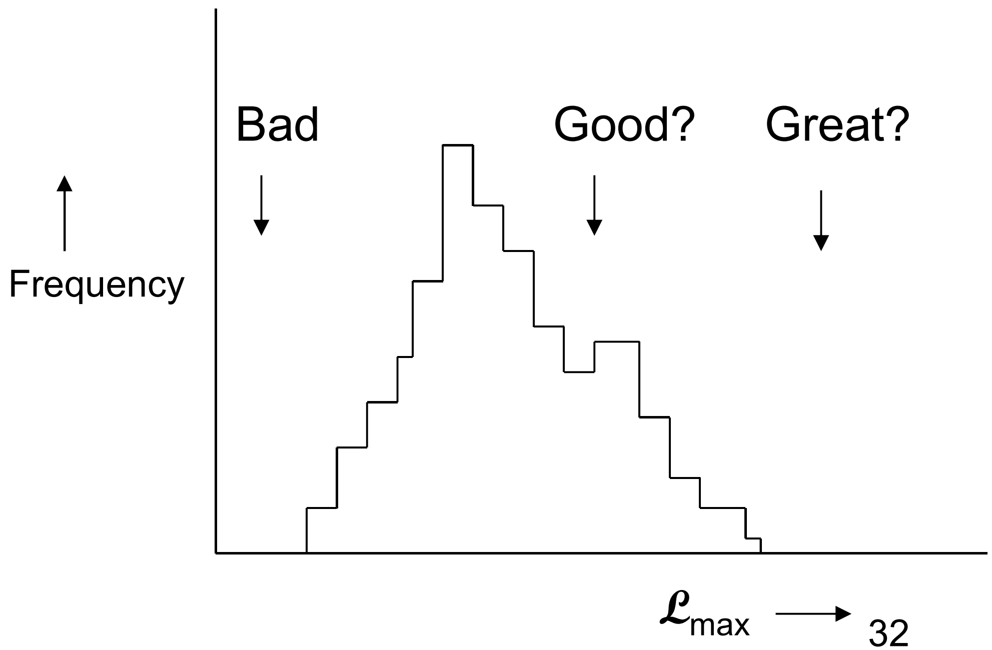
Find params by maximising $\mathcal{L}$

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

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

Monte Carlo distribution

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



Difference between ℒ and pdf

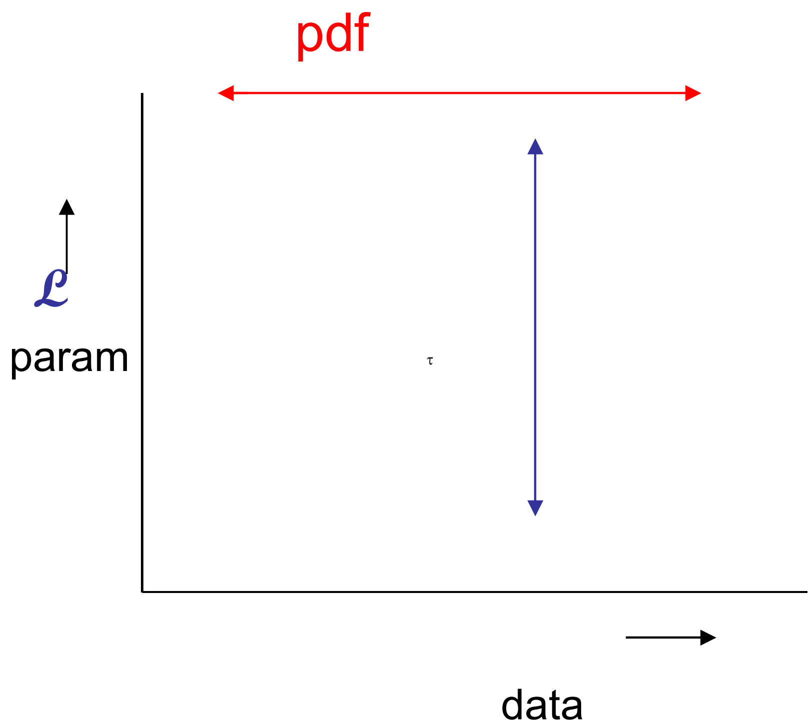
Not necessarily:

ℒ(data,params)

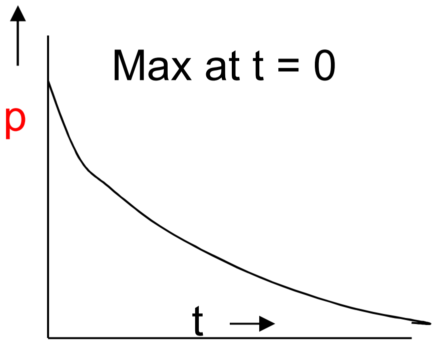
↑ fixed ↑ vary

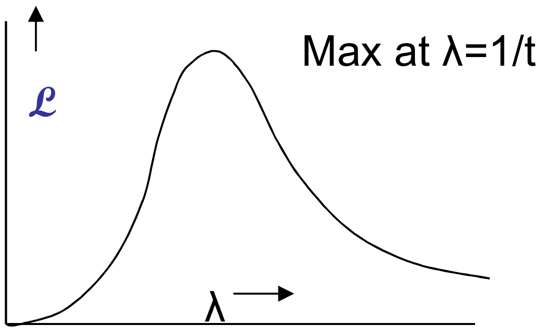
Contrast pdf(data,params)

↑ vary ↑ fixed



e.g. p(λ) = λ exp(-λt)





Example 1

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

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

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

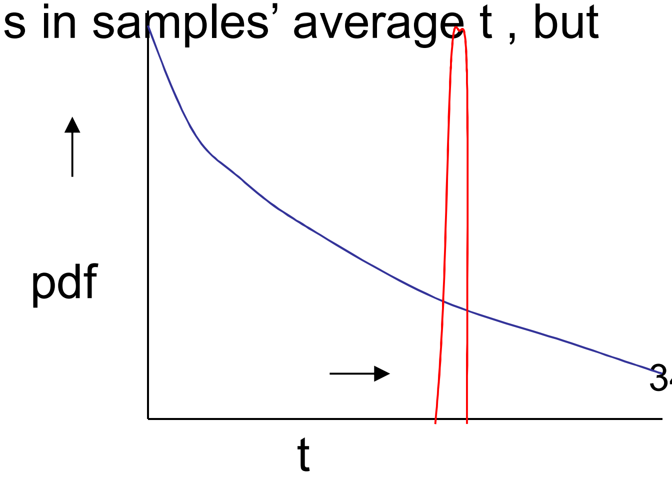
i.e. Depends only on AVERAGE t, but is

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

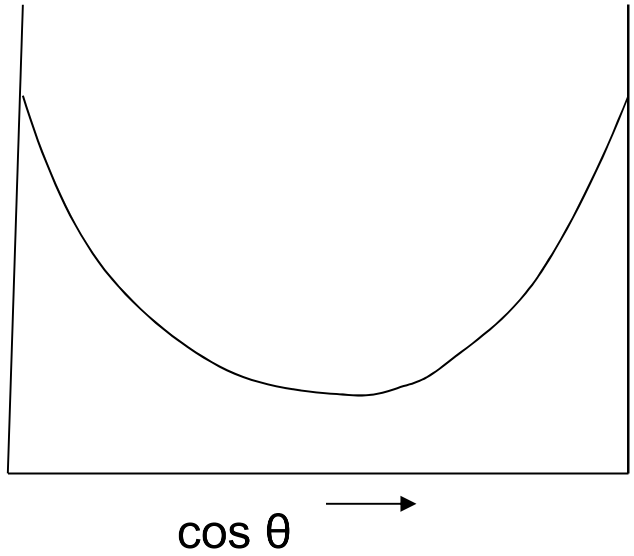


Same average t $\Rightarrow$ same $\mathcal{L}_{max}$

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

Example 3

Fit to Gaussian with variable μ, fixed σ

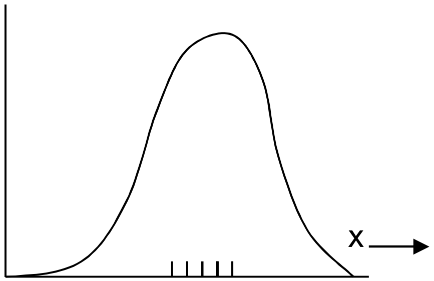
$$\ln \mathcal{L}_{max} = N(-0.5 \ln 2\pi - \ln\sigma) - 0.5\, \Sigma(x_i - x_{av})^2 / \sigma^2$$

constant ~variance(x)

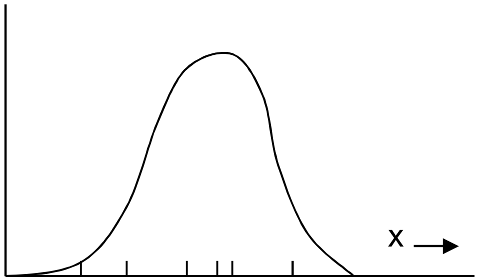
i.e. $\mathcal{L}_{max}$ depends only on variance(x),

which is not relevant for fitting μ ($\mu_{est} = x_{av}$)

Smaller than expected variance(x) results in larger $\mathcal{L}_{max}$



Worse fit, larger $\mathcal{L}_{max}$



Better fit, lower $\mathcal{L}_{max}$

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

Conclusion:

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

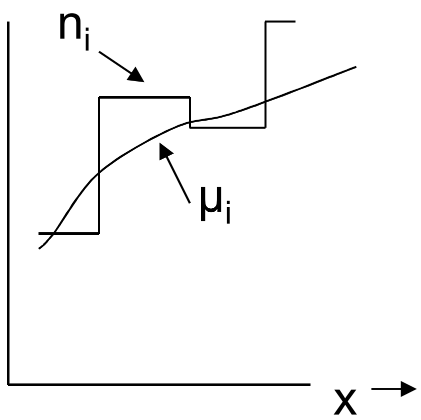
NOT with respect to data

$\mathcal{L}_{max}$ 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



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

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

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

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

$\mathcal{L}_{best}$ 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

$\mathcal{L}$ and pdf

Example 1: Poisson

pdf = Probability density function for observing n, given μ

$$P(n;\mu) = e^{-\mu}\,\mu^n/n!$$

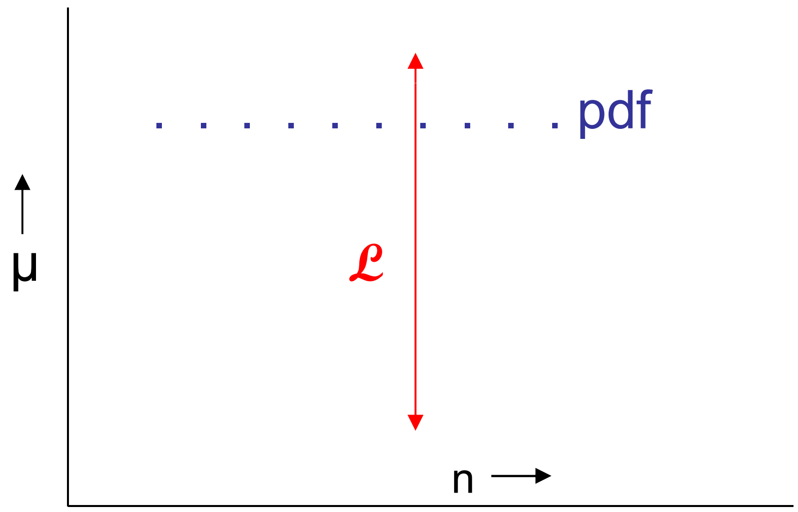
From this, construct $\mathcal{L}$ as

$$\mathcal{L}(\mu;n) = e^{-\mu}\,\mu^n/n!$$

i.e. use same function of μ and n, but

for pdf, μ is fixed, but

for $\mathcal{L}$, n is fixed



N.B. $P(n;\mu)$ exists only at integer non-negative n

$\mathcal{L}(\mu;n)$ exists only as continuous function of non-negative μ

Example 2 Lifetime distribution

pdf $p(t;\lambda) = \lambda e^{-\lambda t}$

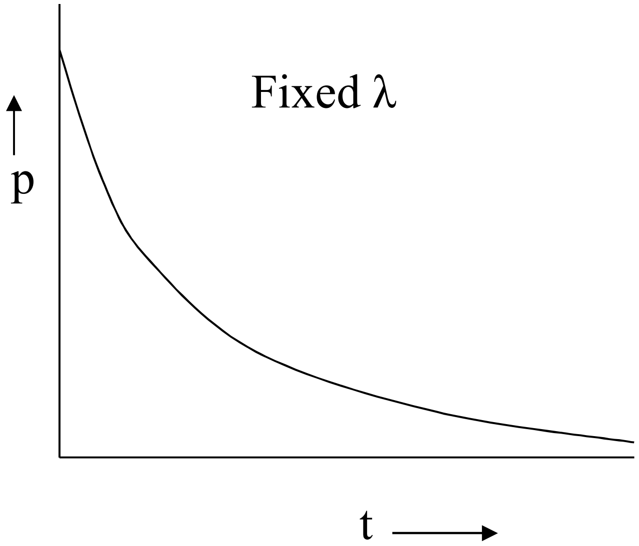
So $\mathcal{L}(\lambda;t) = \lambda e^{-\lambda t}$ (single observed t)

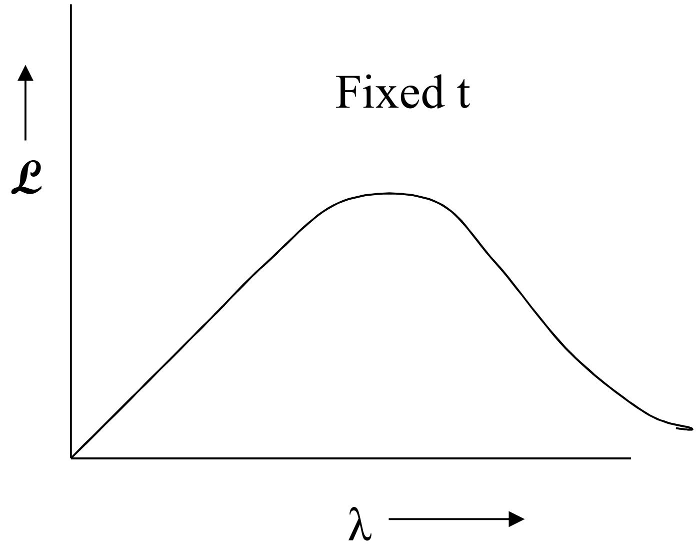
Here both t and λ are continuous

pdf maximises at $t = 0$

$\mathcal{L}$ maximises at $\lambda = t$

N.B. Functional form of $p(t)$ and $\mathcal{L}(\lambda)$ are different





Example 3:   Gaussian

$$\mathrm{pdf}(x;\mu) = \exp\{-(x-\mu)^2/2\sigma^2\} \;/(\sigma\sqrt{2\pi})$$

$$\mathcal{L}(\mu;x) \quad = \exp\{-(x-\mu)^2/2\sigma^2\} \;/(\sigma\sqrt{2\pi})$$

N.B. In this case, same functional form for pdf and $\mathcal{L}$

So if you consider just Gaussians, can be confused between pdf and $\mathcal{L}$

So examples 1 and 2 are useful

Transformation properties of pdf and $\mathcal{L}$

Lifetime example: dn/dt = $\lambda e^{-\lambda t}$

Change observable from t to y = $\sqrt{t}$

$$\frac{dn}{dy} = \frac{dn}{dt}\frac{dt}{dy} = 2y\lambda e^{-\lambda y^2}$$

So (a) pdf changes, BUT

(b) $$\int_{t_0}^{\infty} \frac{dn}{dt} dt = \int_{\sqrt{t_0}}^{\infty} \frac{dn}{dy} dy$$

i.e. corresponding integrals of pdf are
INVARIANT

Now for $\mathcal{L}$ikelihood

When parameter changes from λ to $\tau = 1/\lambda$

(a') $\mathcal{L}$ does not change

dn/dt = $(1/\tau) \exp\{-t/\tau\}$

and so $\mathcal{L}(\tau;t) = \mathcal{L}(\lambda=1/\tau;t)$

because identical numbers occur in evaluations of the two $\mathcal{L}$'s

BUT

(b')

So it is NOT meaningful to integrate $\mathcal{L}$

(However,………)

	pdf(t;λ)	$\mathcal{L}$(λ;t)
Value of function	Changes when observable is transformed	INVARIANT wrt transformation of parameter
Integral of function	INVARIANT wrt transformation of observable	Changes when param is transformed
Conclusion	Max prob density not very sensible	Integrating $\mathcal{L}$ not very sensible

CONCLUSION:

$$\int_{p_l}^{p_u} L\,dp = \alpha$$ NOT recognised statistical procedure

[Metric dependent:

τ range agrees with τ_{pred}

λ range inconsistent with $1/\tau_{pred}$]

BUT

1) Could regard as "black box"

2) Make respectable by $\mathcal{L}$ ⟹ Bayes' posterior

Posterior(λ) ~ $\mathcal{L}$(λ)* Prior(λ) [and Prior(λ) can be constant]

6) BAYESIAN SMEARING OF $\mathcal{L}$

"USE $\ln \mathcal{L}$ FOR $\hat{p}$ + σ_p

SMEAR IT TO INCORPORATE

SYSTEMATIC UNCERTAINTIES"

SCENARIO:

$$n = \text{Poisson}(\mu = s\epsilon + b)$$

PARAM OF INTEREST → s

EFFIC / ACCEPTANCE / $\int \mathcal{L}$ → ϵ

BACKGROUND → b

UNCERTAINTIES MEASURED IN 'SUBSIDIARY' EXPT

$$P(s, \epsilon | n) = \frac{P(n | s, \epsilon)\, \pi(s, \epsilon)}{\iint \ldots\ldots\ldots\ldots \, ds\, d\epsilon}$$

($P(n | s, \epsilon)$ = $\mathcal{L}$)

$$P(s|n) = \int P(s, \epsilon | n)\, d\epsilon$$

$$= \frac{\int \mathcal{L}\, \pi(s)\, \pi(\epsilon)\, d\epsilon}{\iint \ldots\ldots\ldots \, ds\, d\epsilon}$$

e.g. $\pi(s)$ = truncated const. $\pi(\epsilon) \sim e^{-\frac{1}{2}\left(\frac{\epsilon - \epsilon_0}{\sigma}\right)^2}$ [BEWARE]

i.e. SMEAR $\mathcal{L}$ (<u>not</u> $\ln \mathcal{L}$) by "prior" for ϵ

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

$f_A = 1/3$

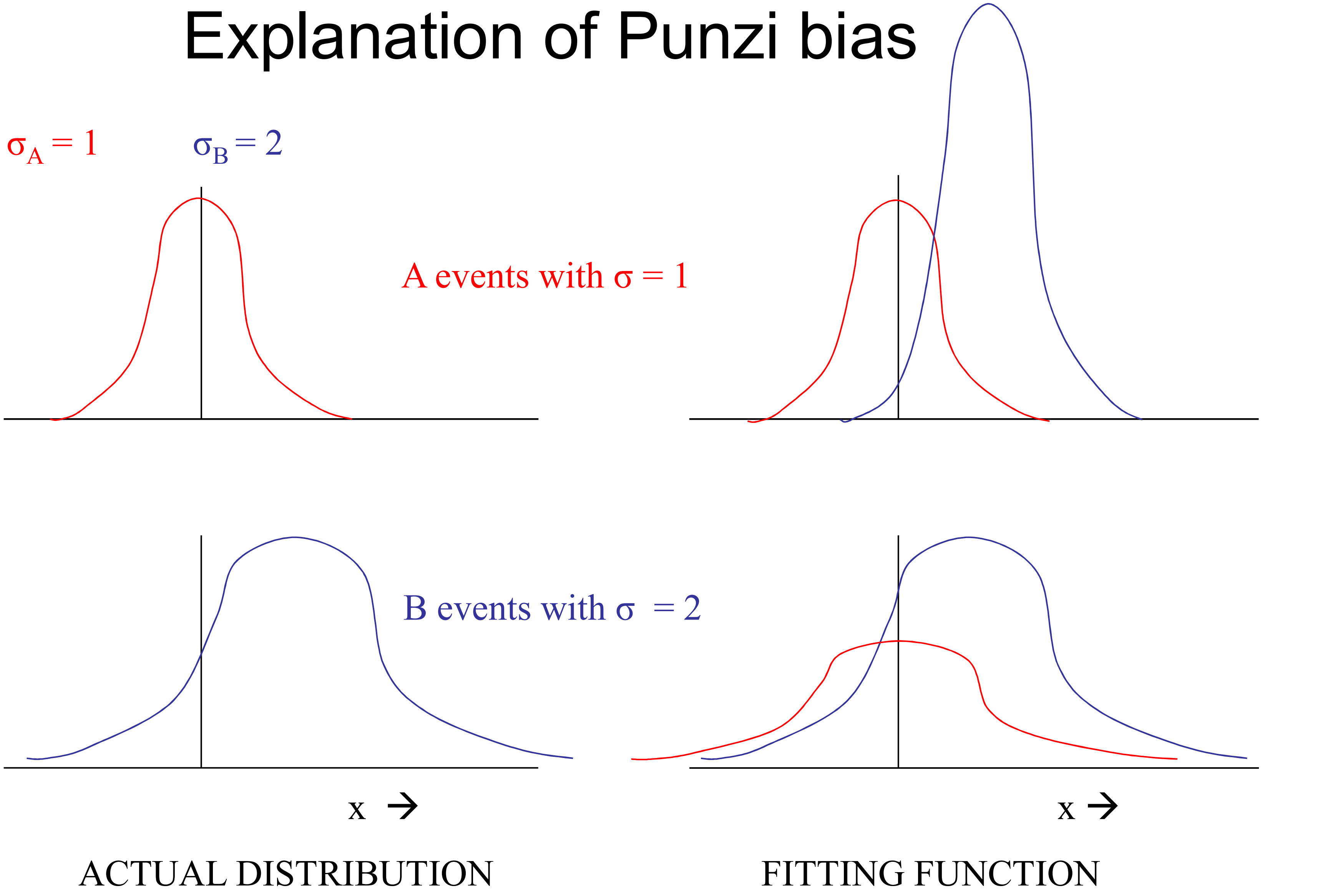
σ_A	σ_B	$\mathcal{L}_{wrong}$ f_A	$\mathcal{L}_{wrong}$ σ_f	$\mathcal{L}_{right}$ f_A	$\mathcal{L}_{right}$ σ_f
1·0	1·0	0·336(3)	0·08	Same	
1·0	1·1	0·374(4)	0·08	0· 333(0)	0
1·0	2·0	0·645(6)	0·12	0·333(0)	0
1 → 2	1.5 →3	0·514(7)	0·14	0·335(2)	0·03
1.0	1 → 2	0.482(9)	0.09	0.333(0)	0

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

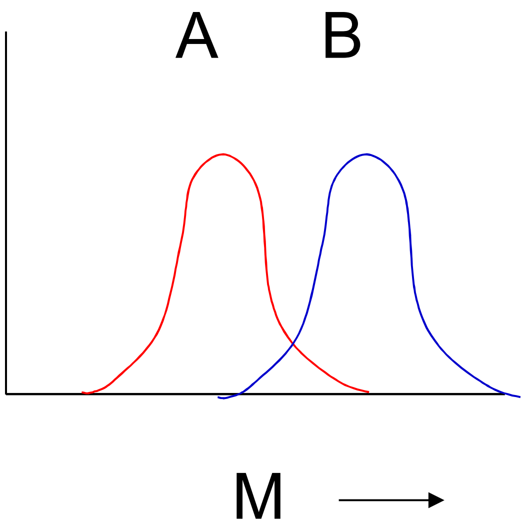


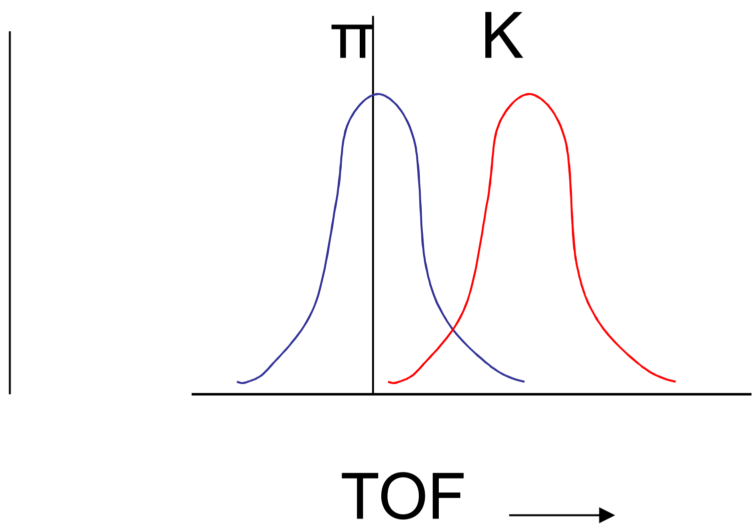
[N_A/N_B variable, but same for A and B events]

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(σ|A) and p(σ|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

What else can we do with $\mathcal{L}$s?

So far mainly parameter determination (also Baker & Cousins' Goodness of Fit with Likelihood ratio)

Other possibilities:

Frequentist approach:

Construction of parameter confidence intervals

Likelihood ratios for comparing Hypotheses

Bayesian approach:

Together with priors → parameter credible intervals;

and Comparing Hypotheses

More in lectures by Olaf Behnke & Glen Cowan

BAYES and FREQUENTISM
The Return of an Old Controversy

Parameter Determination

We need to make a statement about

Parameters, Given Data

The basic difference between the two:

Bayesian : Prob(parameter, given data)
(an anathema to a Frequentist!)

Frequentist : Prob(data, given parameter)
(a likelihood function)

WHAT IS PROBABILITY?

MATHEMATICAL

Formal

Based on Axioms

FREQUENTIST

Ratio of frequencies as n→ infinity

Repeated “identical” trials

Not applicable to **single event** or **physical constant**

BAYESIAN Degree of belief

Can be applied to single event or physical constant

(even though these have unique truth)

Varies from person to person ***

Quantified by “fair bet”

Picture of Bayes

LEGAL PROBABILITY

Picture of Reverend Bayes

Maybe it isn't Bayes?

"Probability that this is actually a picture of Bayes" is not Frequentist probability.

"Probability of Bayes" is Bayesian probability.

Bayesian

$$P(A;B) = \frac{P(B;A) \text{ x } P(A)}{P(B)}$$

Bayes' Theorem

p(param | data) α p(data | param) * p(param)

↑ posterior ↑ likelihood ↑ prior

Problems: p(param) Has particular value

"Degree of belief"

Prior What functional form?

Coverage

$P\ (Data;Theory) \neq P\ (Theory;Data)$

(Example of $P(A;B) \neq P(B;A)$)

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Classical Approach: Neyman Construction

Neyman "confidence interval" avoids pdf for μ

Uses only $P(x;\mu)$

Confidence interval $\mu_1 \rightarrow \mu_2$:

$P(\mu_1 \rightarrow \mu_2$ contains $\mu_t) = \alpha$ True for any μ_t

Varying intervals from ensemble of experiments

fixed

Gives range of μ for which observed value x_0 was "likely" (α)

Contrast Bayes : Degree of belief = α that μ_t is in $\mu_1 \rightarrow \mu_2$

Classical (Neyman) Confidence Intervals

Uses only P(data|theory)

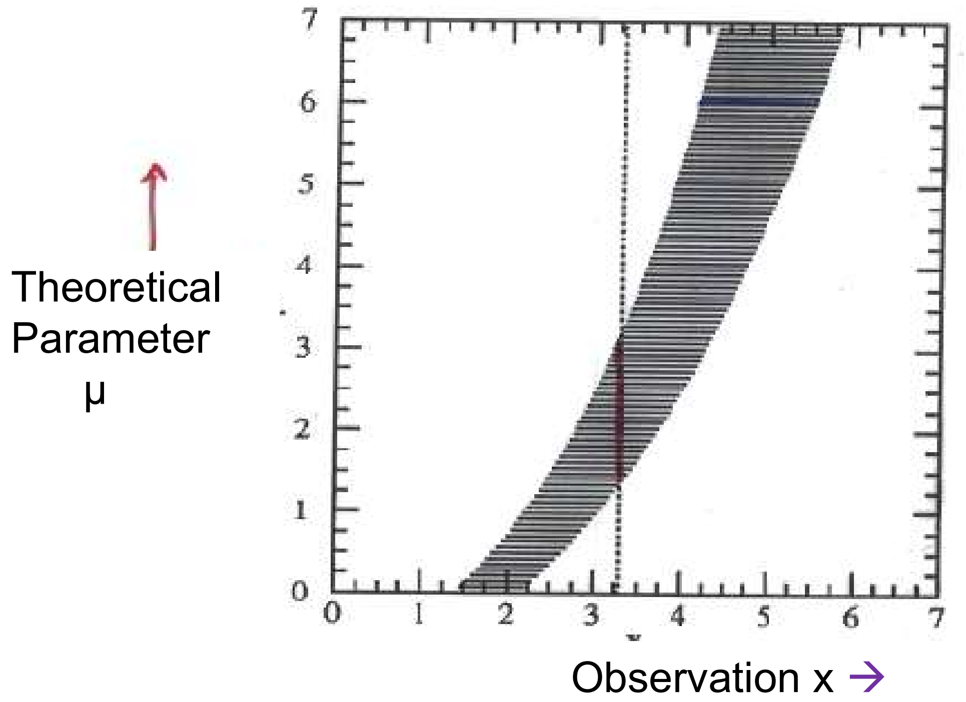


FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2] | \mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0. The confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.

$\mu \geq 0$ No prior for μ

$\mu_l \le \mu \le \mu_u$ at 90% confidence

Frequentist

μ_l and μ_u known, but random

μ unknown, but fixed

Probability statement about μ_l and μ_u

Bayesian

μ_l and μ_u known, and fixed

μ unknown, and random

Probability/credible statement about μ

Conclusions: What you now know

How it works, and how to estimate uncertainties

$\Delta(\ln \mathcal{L}) = 0.5$ rule and coverage

Several Parameters

Commbining Profile $\mathcal{L}$s loses information

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

Intro to Bayes and Frequentism

FINAL MESSAGE

You cannot become an expert on Statistics by just reading books and listening to lectures.
You have to work at it – solve lots of problems, etc.

Best of luck with Statistics, and with your research and enjoy this School!