

χ^2 and Goodness of Fit

Louis Lyons
IC and Oxford

CERN Latin American School

March 2015

Least squares best fit

Resume of straight line

Correlated errors

Errors in x and in y

Goodness of fit with χ^2

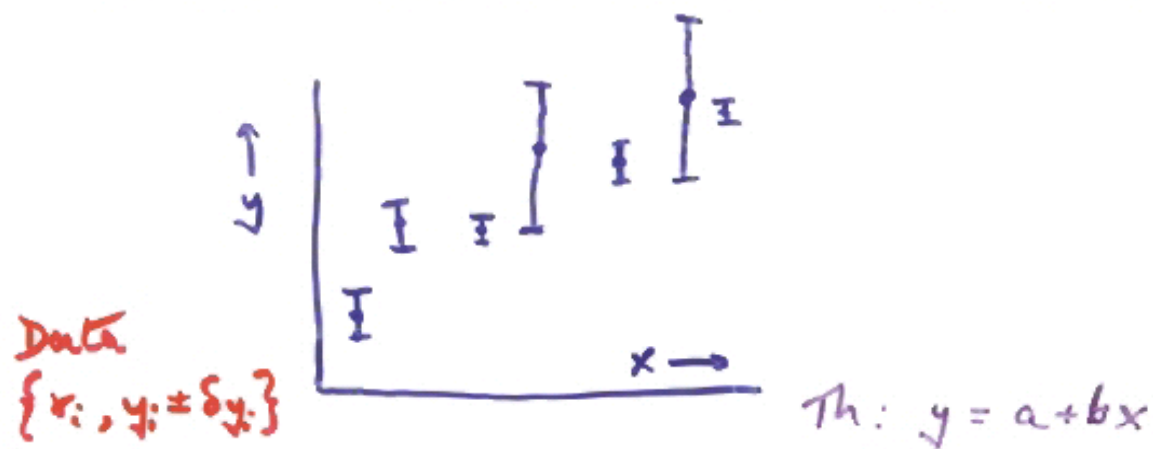
Errors of first and second kind

Kinematic fitting

Toy example

THE paradox

LEAST SQUARES STRAIGHT LINE FITTING



1) DOES IT FIT STRAIGHT LINE?

(HYPOTHESIS TESTING)

2) WHAT ARE GRADIENT + INTERCEPT?

(PARAMETER DETERMINATION)

↑
1st

N.B. 1 CAN BE USED FOR NON - " $a + bx$ "
e.g. $a + b \cos^2 \theta$

N.B. 2. LEAST SQUARES NOT ONLY METHOD

$$S = \sum_i \left(\frac{y_i^{th} - y_i^{obs}}{\sigma_i} \right)^2$$

σ_i SUPPOSED TO BE "ERROR ON TH." *

TAKEN AS "ERROR ON EXPT"

i) Makes algebra simpler

ii) If theory ~ expt, not too different.

IF THEORY (or DATA) O.K.

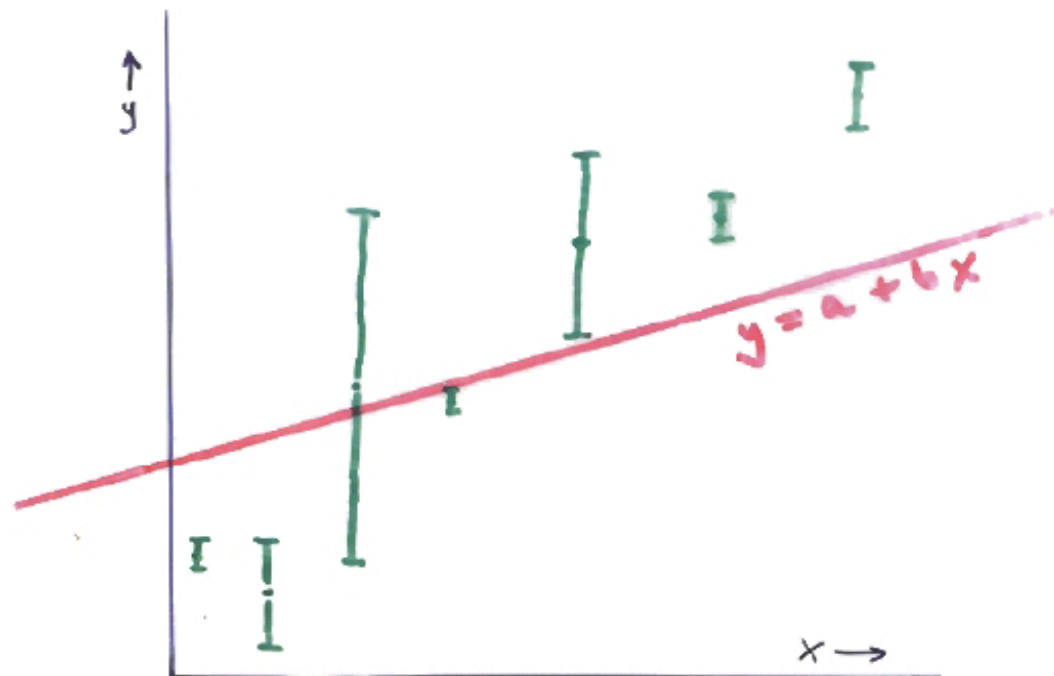
$y^{th} \sim y^{obs} \Rightarrow S$ small

Minimise $S \Rightarrow$ best line

Value of $S_{min} \Rightarrow$ how good fit is.

*

Th	Obs	σ_{th}	σ_{obs}	Chr & S
0.01	1	0.1		100
			1	1



Criterion:

$$S = \sum_i \left(\frac{y_i^{th}(a, b) - y_i^{obs}}{\sigma_i} \right)^2$$

$\xrightarrow{\text{Vert devn}}$
 \uparrow
 An error for each pt.

SIMPLE EXAMPLE OF MINIMISING S

Measurements $\left. \begin{matrix} a_1 \pm \sigma_1 \\ a_2 \pm \sigma_2 \\ \vdots \\ a_i \pm \sigma_i \end{matrix} \right\}$ Best value $\hat{a} \pm \sigma$

Construct $S = \sum \left(\frac{\hat{a} - a_i}{\sigma_i} \right)^2$

Minimise S w.r.t. \hat{a}

$$\frac{1}{2} \frac{\partial S}{\partial \hat{a}} = \sum \frac{\hat{a} - a_i}{\sigma_i^2} = 0$$

$$\hat{a} \sum \frac{1}{\sigma_i^2} = \sum \frac{a_i}{\sigma_i^2} \quad \star$$

Error on \hat{a} given by $\sigma = \left(\frac{1}{2} \frac{\partial^2 S}{\partial \hat{a}^2} \right)^{-1/2}$

$$\frac{\partial^2 S}{\partial \hat{a}^2} = 2 \sum \frac{1}{\sigma_i^2}$$

$$\therefore \frac{1}{\sigma^2} = \sum \frac{1}{\sigma_i^2} \quad \star$$

IN PARABOLIC APX
EQUIV TO
 $S \rightarrow S_{\min} + 1$

Many params

$$\frac{1}{2} \frac{\partial^2 S}{\partial x_i \partial x_j} = \text{INVERSE ERROR MATRIX}$$



Straight Line Fit

$$S = \sum_i \left(\frac{(a + bx_i) - y_i}{\sigma_i} \right)^2$$

i) "Draw" lots of lines $\Rightarrow S$ for each

ii) Minimise S (w.r.t. a & b)

$$\begin{aligned} \frac{1}{2} \frac{\partial S}{\partial a} &= \sum_i \frac{(\overset{\downarrow}{a} + \overset{\downarrow}{b}x_i - y_i)}{\sigma_i^2} = 0 \\ \frac{1}{2} \frac{\partial S}{\partial b} &= \sum_i \frac{(\overset{\downarrow}{a} + \overset{\downarrow}{b}x_i - y_i)x_i}{\sigma_i^2} = 0 \end{aligned} \quad \left. \vphantom{\begin{aligned} \frac{1}{2} \frac{\partial S}{\partial a} &= \sum_i \frac{(\overset{\downarrow}{a} + \overset{\downarrow}{b}x_i - y_i)}{\sigma_i^2} = 0 \\ \frac{1}{2} \frac{\partial S}{\partial b} &= \sum_i \frac{(\overset{\downarrow}{a} + \overset{\downarrow}{b}x_i - y_i)x_i}{\sigma_i^2} = 0 \end{aligned}} \right\} \begin{array}{l} \text{2} \\ \text{SIM. EQNS} \\ \text{FOR 2} \\ \text{UNKNOWN} \\ \text{(\underline{a} \& \underline{b})} \end{array}$$

$$b = \frac{[1][xy] - [x][y]}{[1][x^2] - [x][x]} = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}$$

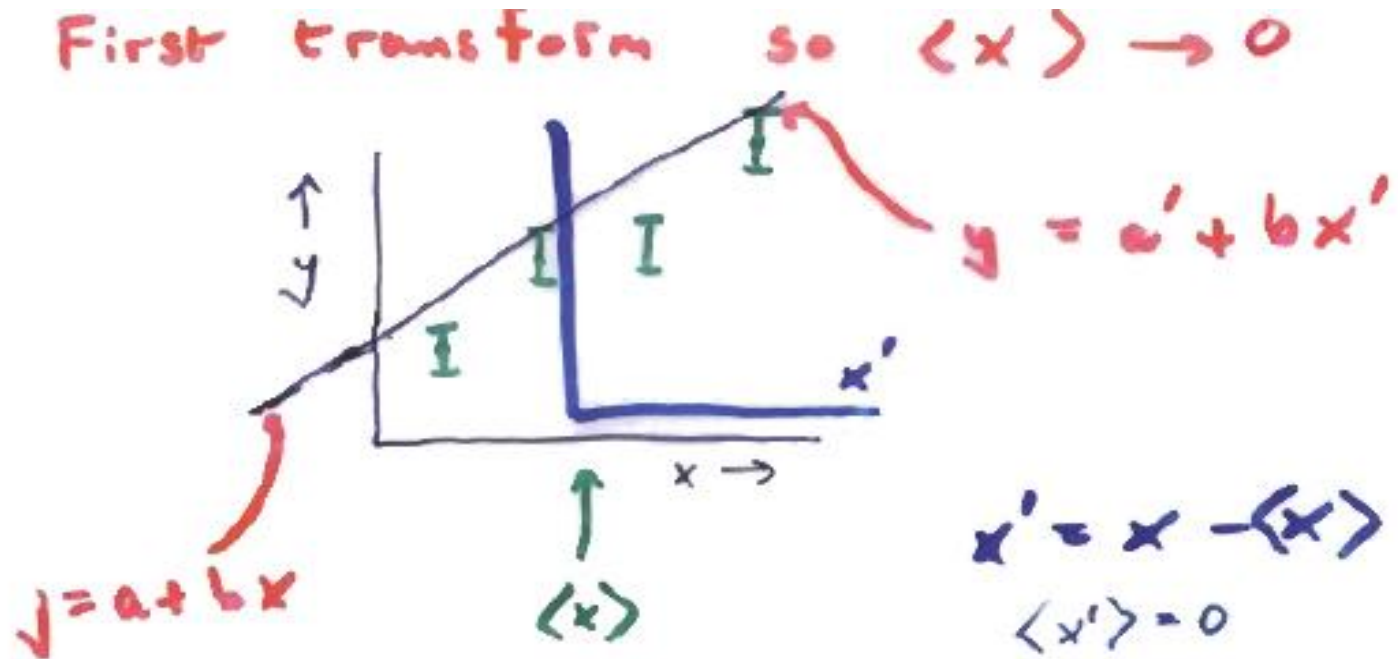
$$\text{where } [f] = \sum \frac{f_i}{\sigma_i^2}$$

$$\text{or } \langle f \rangle = [f]/[1]$$

$$\langle y \rangle = a + b \langle x \rangle \quad \Rightarrow \quad a$$

N.B. L.S.B.F. passes through $(\langle x \rangle, \langle y \rangle)$

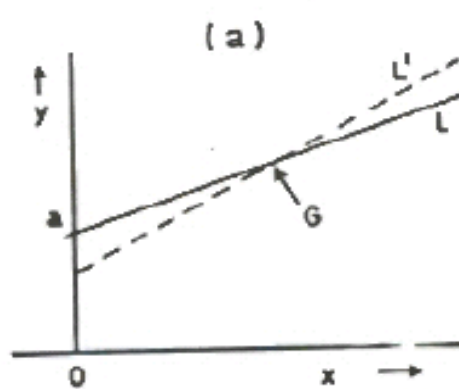
Error on intercept and gradient



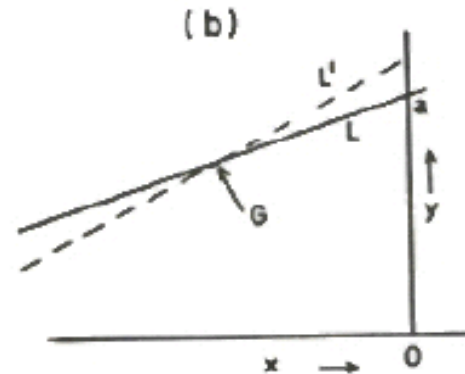
Better to use x' because
error on a' & b are UNCORRELATED
[Cf. Errors on a & b CORRELATED]

That is why track parameters specified at track 'centre'

COVARIANCE (a, b) $\propto -\langle x \rangle$



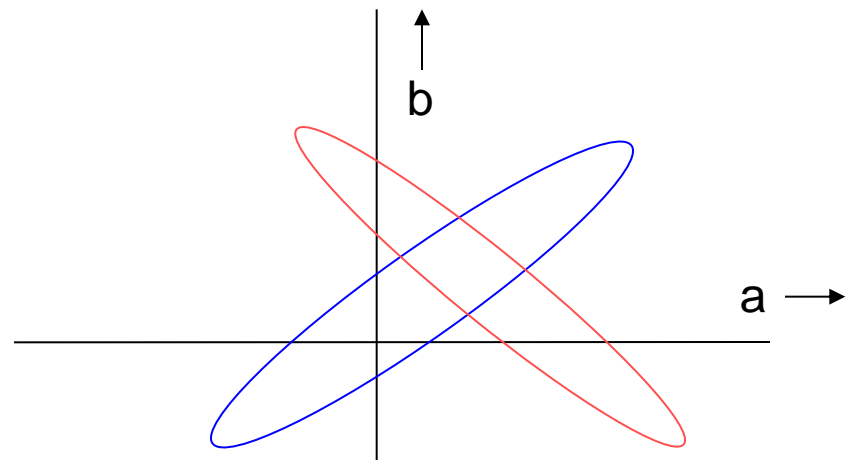
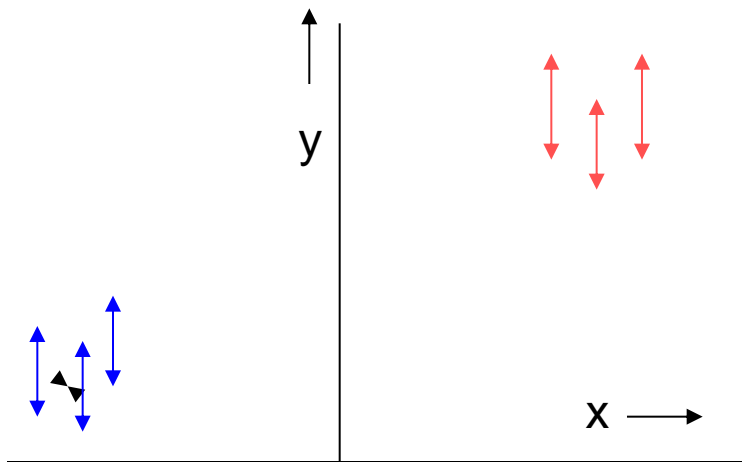
$\langle x \rangle$ pos



$\langle x \rangle$ neg

Fig. 2.4

See Lecture 1



If no errors specified on y_i (!)

ASSUME ALL ERRORS EQUAL
(or similar)

σ CANCELS FROM a & b
e.g. $b = \frac{[1][xy] - [x][y]}{[1][x^2] - [x]^2}$

NEED σ for errors on a' & b

$$S = \frac{1}{\sigma^2} \sum (a + bx_i - y_i)^2 = \chi^2$$
$$\Rightarrow \sigma$$
$$\Rightarrow \sigma(a') \text{ \& } \sigma(b)$$

i.e. USE SCATTER OF POINTS AROUND


STRAIGHT LINE \Rightarrow ERROR ON POINTS

\Rightarrow ERROR ON INTERCEPT + GRADIENT

(cf: Estimate σ from scatter of repeated measurements)

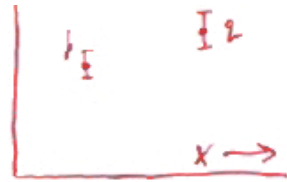
N.B. CANNOT TEST WHETHER DATA IS CONSISTENT
WITH THEORY

Summary of straight line fitting

- Plot data
 - Bad points
 - Estimate a and b (and errors)
- a and b from formula
- Errors on a' and b
- Cf calculated values with estimated
- Determine S_{\min} (using a and b)
- $v = n - p$
- Look up in χ^2 tables 
- If probability too small, **IGNORE RESULTS**
- If probability a “bit” small, scale errors?

 Asymptotically

Measurements with correlated errors e.g. systematics?



Start with 2 uncorrelated measurements

$$S = \frac{(1 - 1_{pr})^2}{\sigma_1^2} + \frac{(2 - 2_{pr})^2}{\sigma_2^2} \quad \#$$

Introduce correlations by

$$\begin{aligned} p &= r \cos \theta - s \sin \theta \\ q &= r \sin \theta + s \cos \theta \end{aligned}$$

NOT ROTN
in x-y SPACE

Write σ_p σ_q (+ $\text{cov}(p, q) = 0$) in terms of σ_r^2 σ_s^2 + $\text{cov}(r, s)$

$$\Rightarrow S = \frac{1}{\sigma_r^2 \sigma_s^2 - \text{cov}(r, s)} \left[\sigma_s^2 (r - r_{pr})^2 + \sigma_r^2 (s - s_{pr})^2 - 2 \text{cov}(r, s) (r - r_{pr})(s - s_{pr}) \right]$$

Inv. est
matrix
element

$$= H_{11} (r - r_{pr})^2 + H_{22} (s - s_{pr})^2 + 2 H_{12} (r - r_{pr})(s - s_{pr})$$

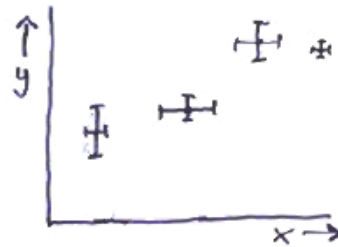
$$\text{where } H^{-1} = \begin{pmatrix} \sigma_r^2 & \text{cov} \\ \text{cov} & \sigma_s^2 \end{pmatrix} \leftarrow \text{Error matrix}$$

Reduces to standard formula in absence of correlus

$$\text{In general : } S = \sum_{ij} \tilde{\Delta}_i H_{ij} \Delta_j$$

$$\text{where } \Delta_j = (\text{observed} - \text{pred.})_j$$

STRAIGHT LINE: Errors on x and on y

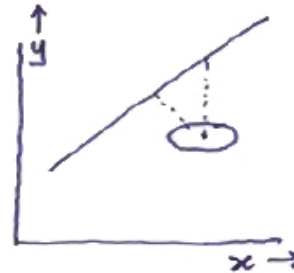


For simplicity,

assume x, y errors uncorrelated

Previously, contribution to S was

$$\left(\frac{y_i - y_i(\text{fit})}{\sigma_i} \right)^2$$



Now replace by

$$\text{Min} \left[\frac{\text{Distance of any point on line, to data point}^2}{\text{Radius of error ellipse in that dirn}} \right]^2$$

i.e. Min of error ellipse function

$$\frac{(x - x_i)^2}{\sigma_{x_i}^2} + \frac{(y - y_i)^2}{\sigma_{y_i}^2} = \frac{(y_i - a - b x_i)^2}{\sigma_{y_i}^2 + b^2 \sigma_{x_i}^2}$$

Best line by minimising $S = \sum \frac{(y_i - a - b x_i)^2}{\sigma_{y_i}^2 + b^2 \sigma_{x_i}^2}$

Errors as usual from $\frac{\partial^2 S}{\partial a^2}$ etc

Analytic soln if all σ_{x_i} same, & also σ_{y_i}

Comments on Least Squares method

1) Need to bin

Beware of too few events/bin

2) Extends to n dimensions



but needs lots of events for n larger than 2 or 3

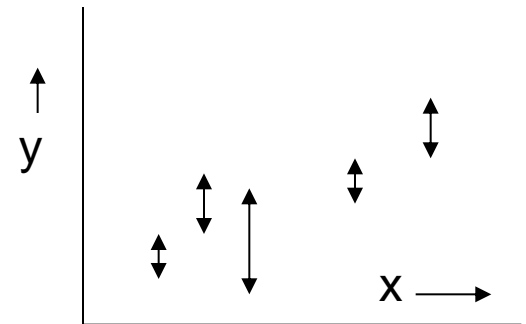
3) No problem with correlated errors

4) Can calculate S_{\min} “on line” i.e. single pass through data

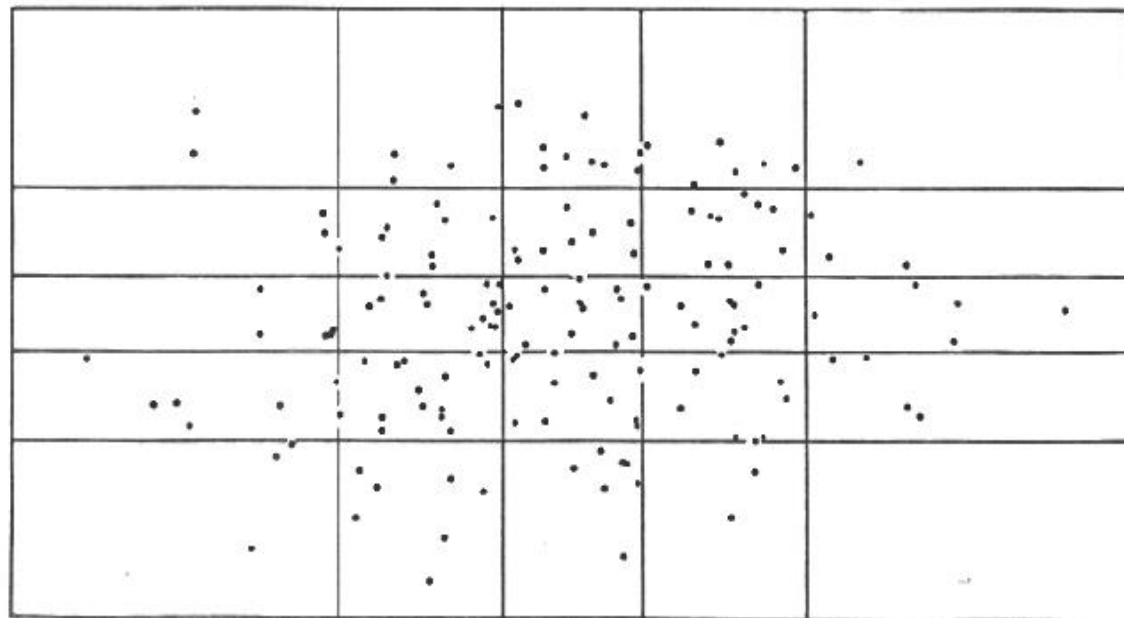
$$\Sigma (y_i - a - bx_i)^2 / \sigma^2 = [y_i^2] - b [x_i y_i] - a [y_i]$$

5) For theory linear in params, analytic solution

6) Hypothesis testing ★ ★ ★



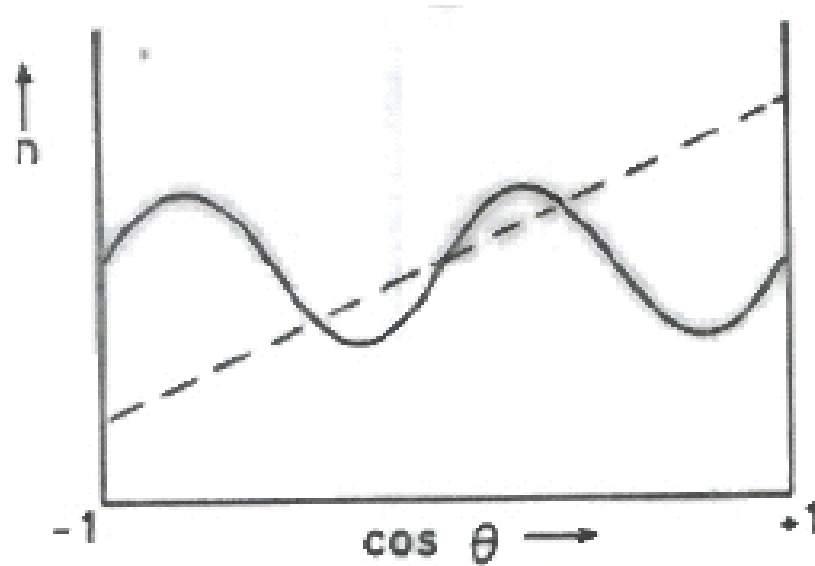
	Individual events (e.g. in $\cos \theta$)	$y_i \pm \sigma_i$ v x_i (e.g. stars)
1) Need to bin?	Yes	No need
4) χ^2 on line	First histogram	Yes



	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
Error estimate	Observed spread, or analytic	$\left\{ - \frac{\partial^2 l}{\partial p_i \partial p_j} \right\}^{-1/2}$	$\left\{ \frac{\partial^2 S}{2 \partial p_i \partial p_j} \right\}^{-1/2}$
Main feature	Easy	Best	Goodness of Fit

‘Goodness of Fit’ by parameter testing?

$$1 + (b/a) \cos^2 \theta \quad \text{Is } b/a = 0 ?$$



‘Distribution testing’ is better

Goodness of Fit: χ^2 test

- 1) Construct S and minimise wrt free parameters
- 2) Determine ν = no. of degrees of freedom

$$\nu = n - p$$

n = no. of data points

p = no. of FREE parameters

- 3) Look up probability that, for ν degrees of freedom,
 $\chi^2 \geq S_{\min}$

Works ASYMPTOTICALLY, otherwise use MC

[Assumes y_i are GAUSSIAN distributed with mean y_i^{th}
and variance σ_i^2]

$$\overline{\chi^2} = \nu$$

$$\sigma^2(\chi^2) = 2\nu$$

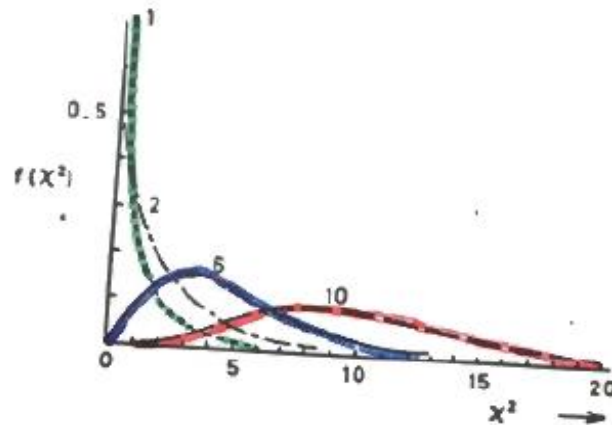


Fig. 2.6

$$\therefore S_{\min} \gtrsim \nu + 3\sqrt{2\nu}$$

is LARGE

e.g. $S_{\min} = 2200$ for $\nu = 2000$?

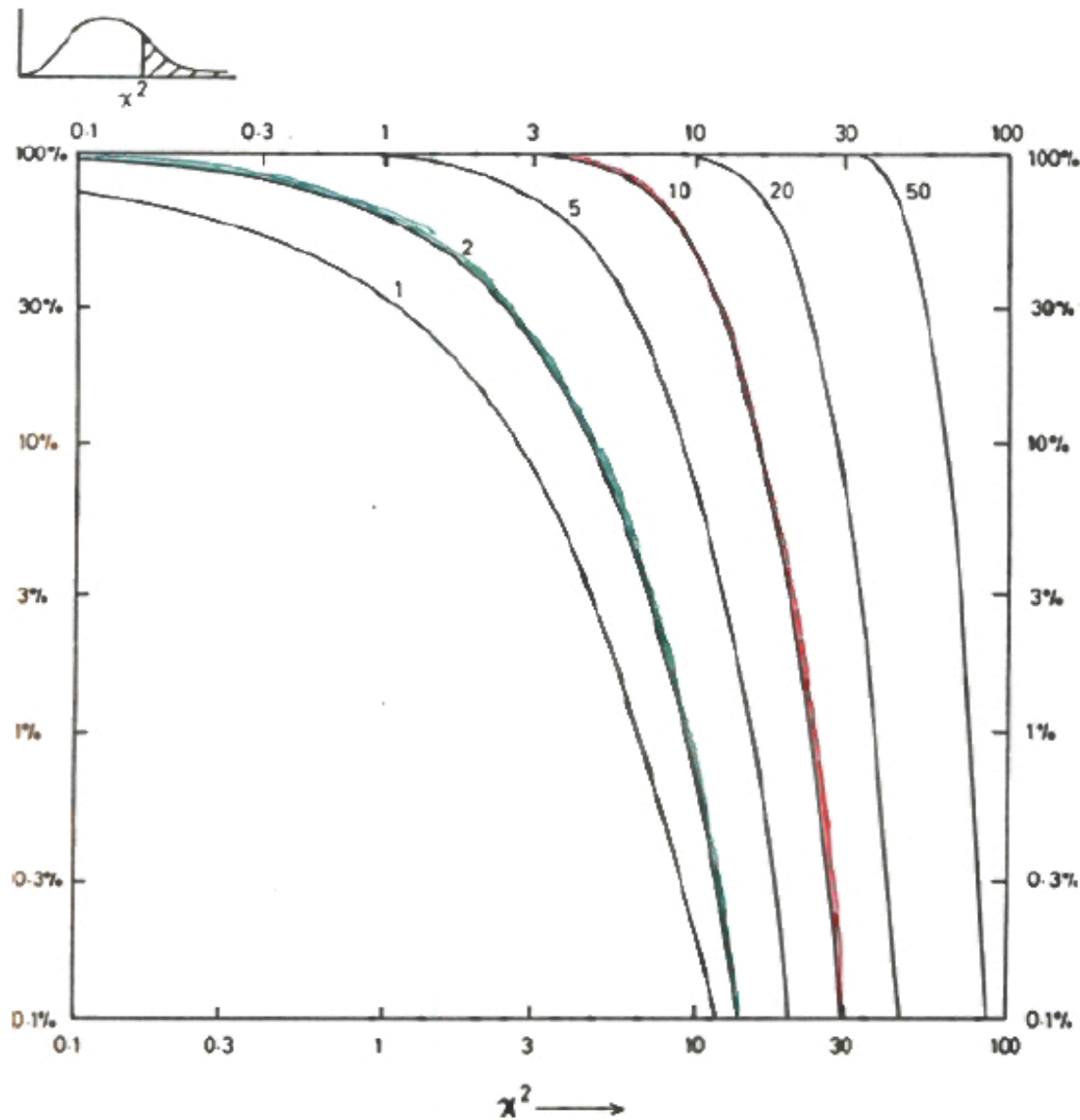


Fig. 2.7

CF: Area in Tails
of Gaussian

χ^2 with ν degrees of freedom?

$\nu = \text{data} - \text{free parameters} ?$

Why asymptotic (apart from Poisson \rightarrow Gaussian) ?

a) Fit flatish histogram with

$$y = N \{ 1 + 10^{-6} \cos(x - \mathbf{x}_0) \} \quad \mathbf{x}_0 = \text{free param}$$

b) Neutrino oscillations: almost degenerate parameters

$$\begin{array}{ll} y \sim 1 - \mathbf{A} \sin^2(1.27 \mathbf{\Delta m^2} L/E) & 2 \text{ parameters} \\ \xrightarrow{\text{Small } \mathbf{\Delta m^2}} 1 - \mathbf{A} (1.27 \mathbf{\Delta m^2} L/E)^2 & 1 \text{ parameter} \end{array}$$

Goodness of Fit

χ^2 : Very general
Needs binning
Not sensitive to sign of dev'n.



Run test

Kolmogorov - Smirnov

etc



See: Aslam + Zech, Durham 1999
Statistics Conf (2002)

Maria Grazia Pin's group in Genoa

Goodness of Fit: Kolmogorov-Smirnov

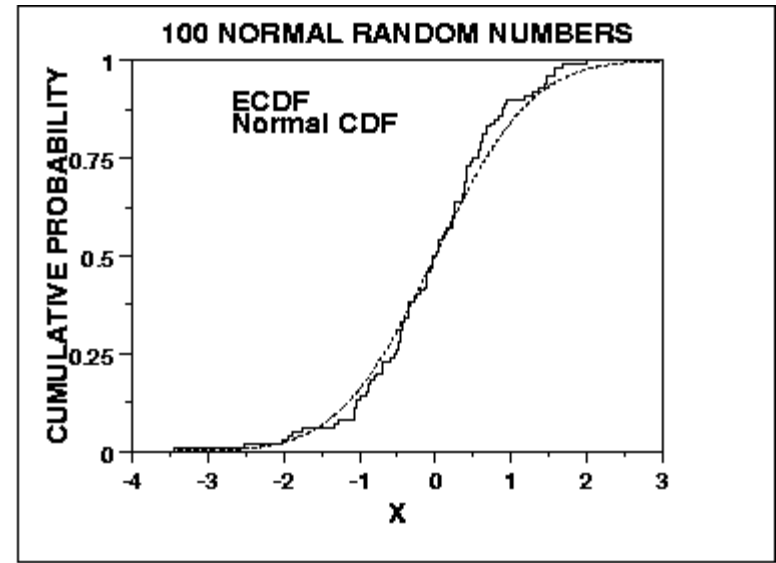
Compares data and model cumulative plots
Uses largest discrepancy between dists.
Model can be analytic or MC sample

Uses individual data points

Not so sensitive to deviations in tails
(so variants of K-S exist)

Not readily extendible to more dimensions

Distribution-free conversion to p; depends on n
(but not when free parameters involved – needs MC)



Goodness of fit: 'Energy' test

Assign +ve charge to data \star ; -ve charge to M.C. \star

Calculate 'electrostatic energy E ' of charges

If distributions agree, $E \sim 0$

If distributions don't overlap, E is positive

Assess significance of magnitude of E by MC

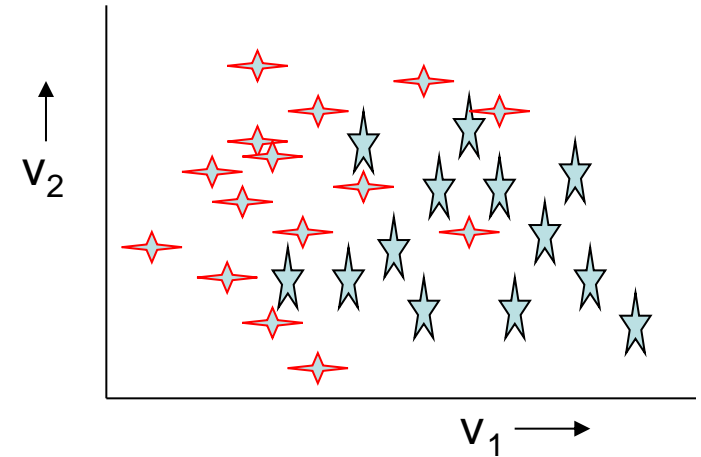
N.B.

- 1) Works in many dimensions
- 2) Needs metric for each variable (make variances similar?)
- 3) $E \sim \sum q_i q_j f(\Delta r = |r_i - r_j|)$, $f = 1/(\Delta r + \epsilon)$ or $-\ln(\Delta r + \epsilon)$

Performance insensitive to choice of small ϵ

See Aslan and Zech's paper at:

<http://www.ippp.dur.ac.uk/Workshops/02/statistics/program.shtml>



Wrong Decisions

Error of First Kind

Reject H_0 when true

Should happen x% of tests

Errors of Second Kind

Accept H_0 when something else is true

Frequency depends on

i) How similar other hypotheses are

e.g. $H_0 = \mu$

Alternatives are: e π K p

ii) Relative frequencies: 10^{-4} 10^{-4} 1 0.1 0.1

Aim for maximum efficiency ← Low error of 1st kind

maximum purity ← Low error of 2nd kind

As χ^2 cut tightens, efficiency ↑ and purity ↓

Choose compromise

How serious are errors of 1st and 2nd kind?

1) Result of experiment

e.g Is spin of resonance = 2?

Get answer WRONG

Where to set cut?

Small cut  Reject when correct

Large cut  Never reject anything

Depends on nature of H0 e.g.

Does answer agree with previous expt?

Is expt consistent with special relativity?

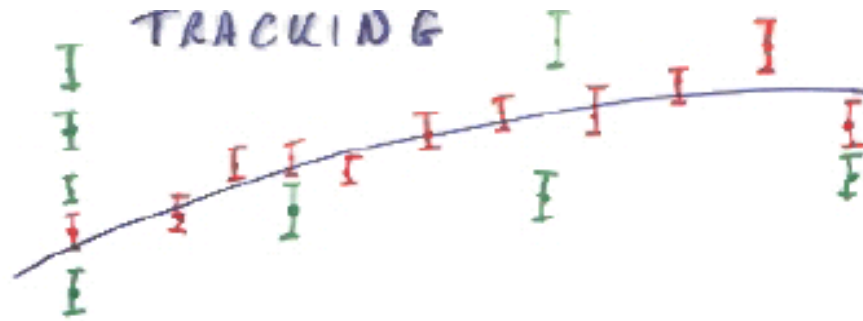
2) Class selector e.g. b-quark / galaxy type / γ -induced cosmic shower

Error of 1st kind: Loss of efficiency

Error of 2nd kind: More background

Usually easier to allow for 1st than for 2nd

3) Track finding



Goodness of Fit: = Pattern Recognition

= Find hits that belong to track

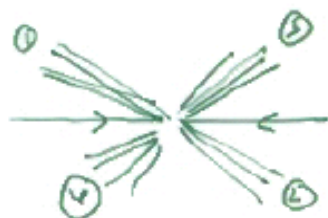
Parameter Determination = Estimate track parameters
(and error matrix)

KINEMATIC FITTING

Test whether observed event consistent with specified reaction

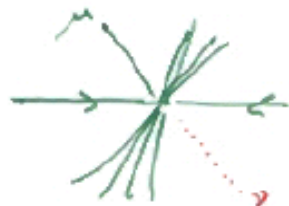


$$\bar{p}p \rightarrow \bar{p}p \pi^+ \pi^-?$$



$$e^+e^- \rightarrow W^+W^- \rightarrow j_1 j_2 j_3 j_4$$

M_W , jet pairings



$$e^+e^- \rightarrow W^+W^- \rightarrow \mu \nu$$

$j_1 j_2$



$$\Lambda \rightarrow p \pi^- \text{ from prodn vertex}$$




$$p + \pi^- \text{ interact}$$

$$\& \Lambda \rightarrow p \pi^- \text{ from prodn vert.}$$

Kinematic Fitting: Why do it?

- 1) CHECK WHETHER EVENT CONSISTENT WITH
HYPOTHESIS [HYPOTHESIS TESTING]
- 2) CAN CALCULATE MISSING VARIABLES [PARAM
DET.N.]
- 3) GOOD TO HAVE TRACKS CONSERVING E-P [P.D.]
- 4) IMPROVES ERRORS [P.D.]

Kinematic Fitting: Why do it?

- 1) CHECK WHETHER EVENT CONSISTENT WITH HYPOTHESIS
[HYPOTHESIS TESTING]
Use S_{min} & No of constraints degrees of freedom
- 2) CAN CALCULATE MISSING VARIABLES [PARAM DETERM.]
e.g. $|P|$ for straight / short track / incoming ν
3 momentum of n, ν, \dots
- 3) GOOD TO HAVE TRACKS CONSERVING $E=P$ [P.D.]
e.g. identical values for resonance mass from prodn or from decay

- 4) IMPROVES ERRORS [P.D.]
Example of
"Adding Theoretical Input can improve error"

Measured variables

$$p\bar{p} \rightarrow p\bar{p}\pi^+\pi^- \quad \star$$

4 momenta of each track

(ie. 3 momenta + assumed/measured track identity)

Then test hypothesis:

Observed event = example of reaction \star

Tested by:

Observed tracks should conserve $E-p$

Can tracks be "wiggled a bit" in order to do so?

$$\text{ie. } S_{\min} = \sum_{\substack{4 \text{ tracks} \\ x \in \mathbb{R}^4}} \left(\frac{v_i^{\text{fitted}} - v_i^{\text{meas}}}{\sigma_i} \right)^2 \quad \leftarrow \text{if uncorr.}$$

Otherwise use Inv. Err. Matrix

where v_i^{fitted} conserve 4-momenta

i.e. Minimisation subject to constraint

(involves Lagrange multipliers)

KINEMATIC FITTING

Angles of triangle: $\theta_1 + \theta_2 + \theta_3 = 180$

	θ_1	θ_2	θ_3	
Measured	50	60	73 ± 1	Sum = 183
Fitted	49	59	72	180

$$\chi^2 = (50-49)^2/1^2 + 1 + 1 = 3$$

$$\text{Prob} \{ \chi^2_1 > 3 \} = 8.3\%$$

ALTERNATIVELY:

Sum = 183 ± 1.7 , while expect 180

$$\text{Prob}\{\text{Gaussian 2-tail area beyond } 1.73 \sigma\} = 8.3\%$$

Toy example of Kinematic Fit



+ constraints:

- 1) Coplanar
- 2) p_1 at θ_1
- 3) p_2 at θ_2
- 4) θ_1 or θ_2

\Leftarrow Non-relativistic equal mass
elastic scatter : $\theta_1 + \theta_2 = \pi/2$

Measured $\theta_1^m \pm \sigma$ $\theta_2^m \pm \sigma$
Fitted θ_1 θ_2

Minimise $S(\theta_1, \theta_2) = \frac{(\theta_1 - \theta_1^m)^2}{\sigma^2} + \frac{(\theta_2 - \theta_2^m)^2}{\sigma^2}$

subject to $C(\theta_1, \theta_2) = \theta_1 + \theta_2 - \pi/2 = 0$

Lagrange : $\frac{\partial S}{\partial \theta_1} + \lambda \frac{\partial C}{\partial \theta_1} = \frac{\partial S}{\partial \theta_2} + \lambda \frac{\partial C}{\partial \theta_2} = 0$

\Rightarrow 3 eqns for θ_1 θ_2 λ

Eqs simple to solve because

$C(\theta_1, \theta_2)$ linear in θ_1, θ_2

$$\Rightarrow \theta_1 = \theta_1^m + \frac{1}{2}(\pi/2 - \theta_1^m - \theta_2^m)$$

$$\theta_2 = \theta_2^m + \frac{1}{2}(\pi/2 - \theta_1^m - \theta_2^m)$$

$$\sigma(\theta_1) = \sigma(\theta_2) = \sigma/\sqrt{2} \quad \star$$

i.e. KINEMATIC FIT \Rightarrow

REDUCED ERRORS

PARADOX

Histogram with 100 bins

Fit with 1 parameter

S_{\min} : χ^2 with NDF = 99 (Expected $\chi^2 = 99 \pm 14$)

For our data, $S_{\min}(p_0) = 90$

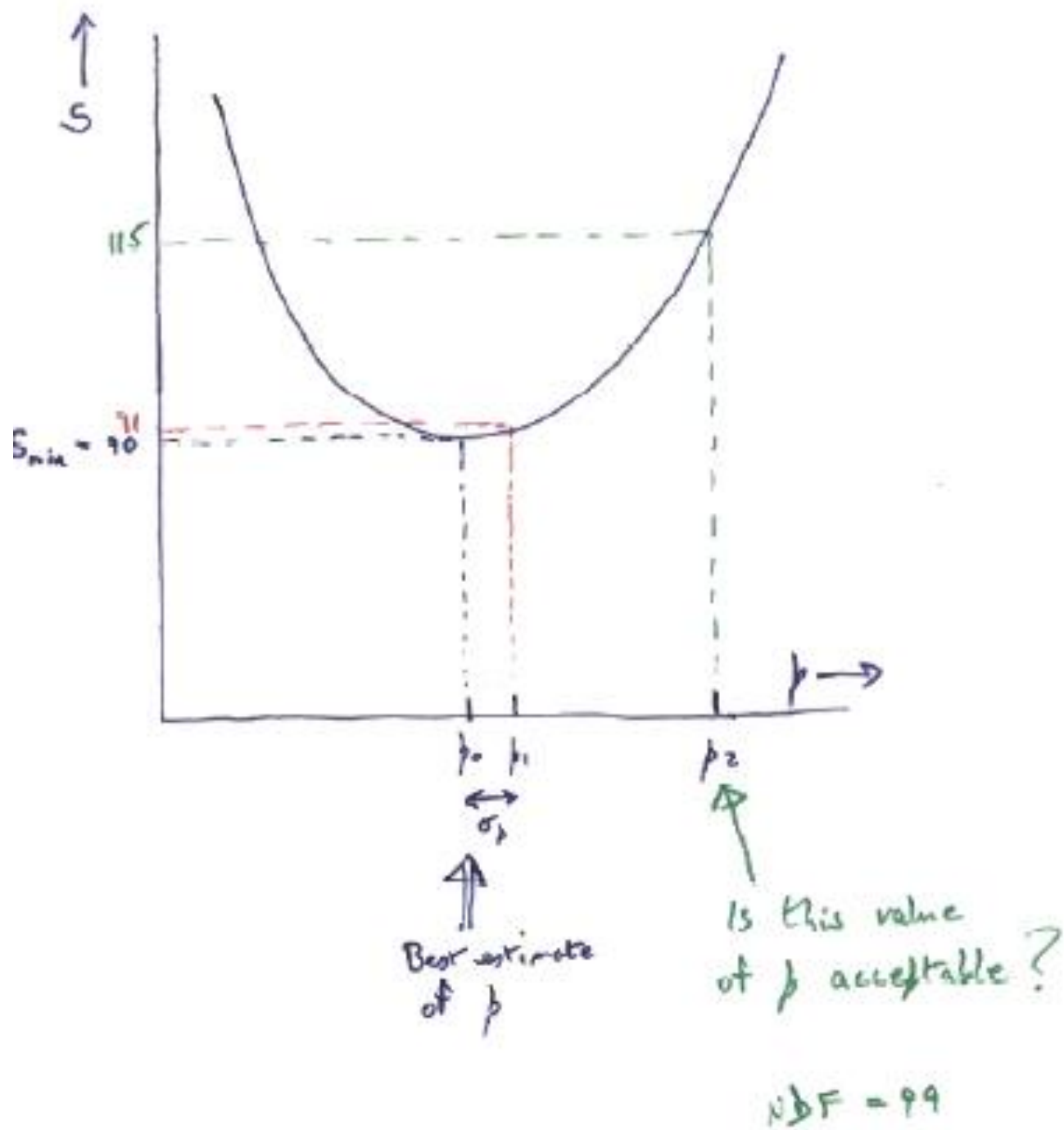
Is p_2 acceptable if $S(p_2) = 115$?

1) YES. Very acceptable χ^2 probability

2) NO. σ_p from $S(p_0 + \sigma_p) = S_{\min} + 1 = 91$

But $S(p_2) - S(p_0) = 25$

So p_2 is 5σ away from best value



Next time:
Discovery and p-values

LHC moves us from era of
‘Upper Limits’ to that of
DISCOVERIES!

Do's and Dont's with *Likelihoods*

Louis Lyons
IC and Oxford
CMS

CERN Latin American School
March 2015

Topics

What it is

How it works: Resonance

Error estimates

Detailed example: Lifetime

Several Parameters

Extended maximum \mathcal{L}

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

Simple example: 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

$$L(\beta) = \prod y_i$$

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

Best estimate of β is that which maximises L

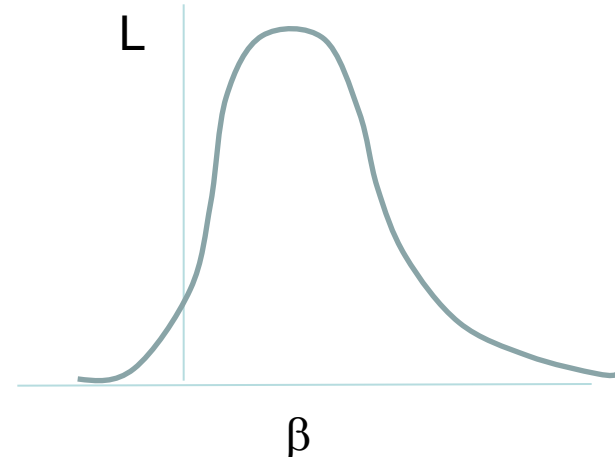
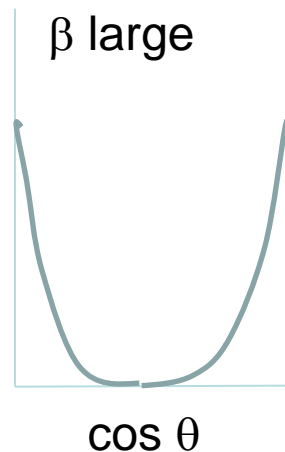
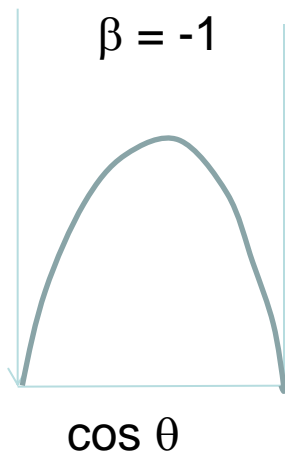
Values of β for which L is very small are ruled out

Precision of estimate for β comes from width of L distribution

CRUCIAL to normalise y

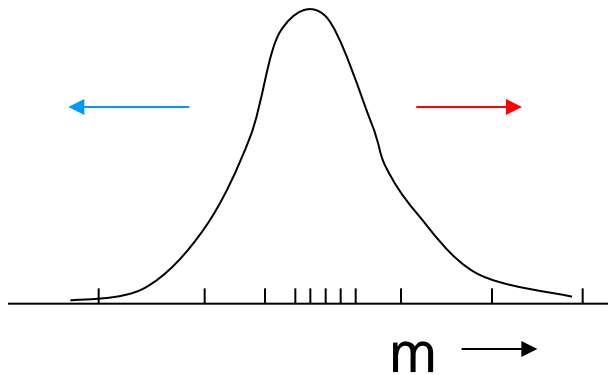
$$N = 1/\{2(1 + \beta/3)\}$$

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

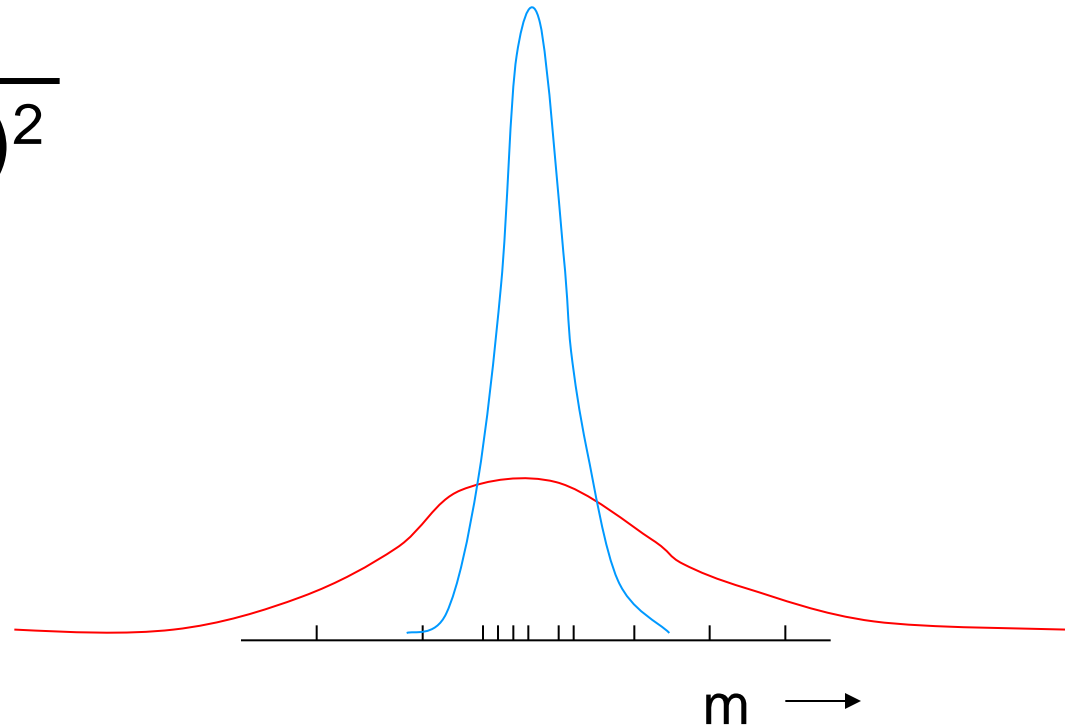


How it works: Resonance

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



Vary M_0



Vary Γ

Conventional to consider

$$l = \ln(\mathcal{L}) = \sum \ln y_i$$

For large N , $\mathcal{L} \rightarrow$ Gaussian

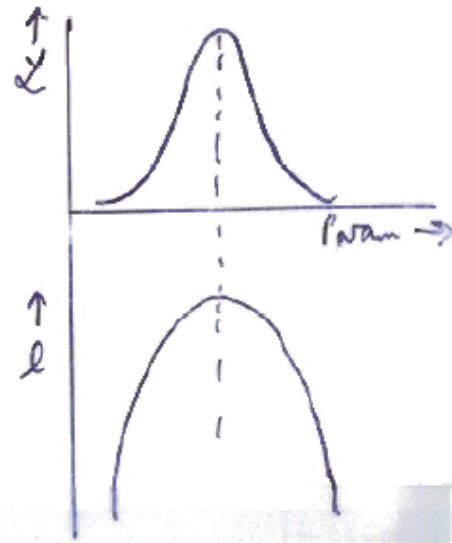
"Proof"

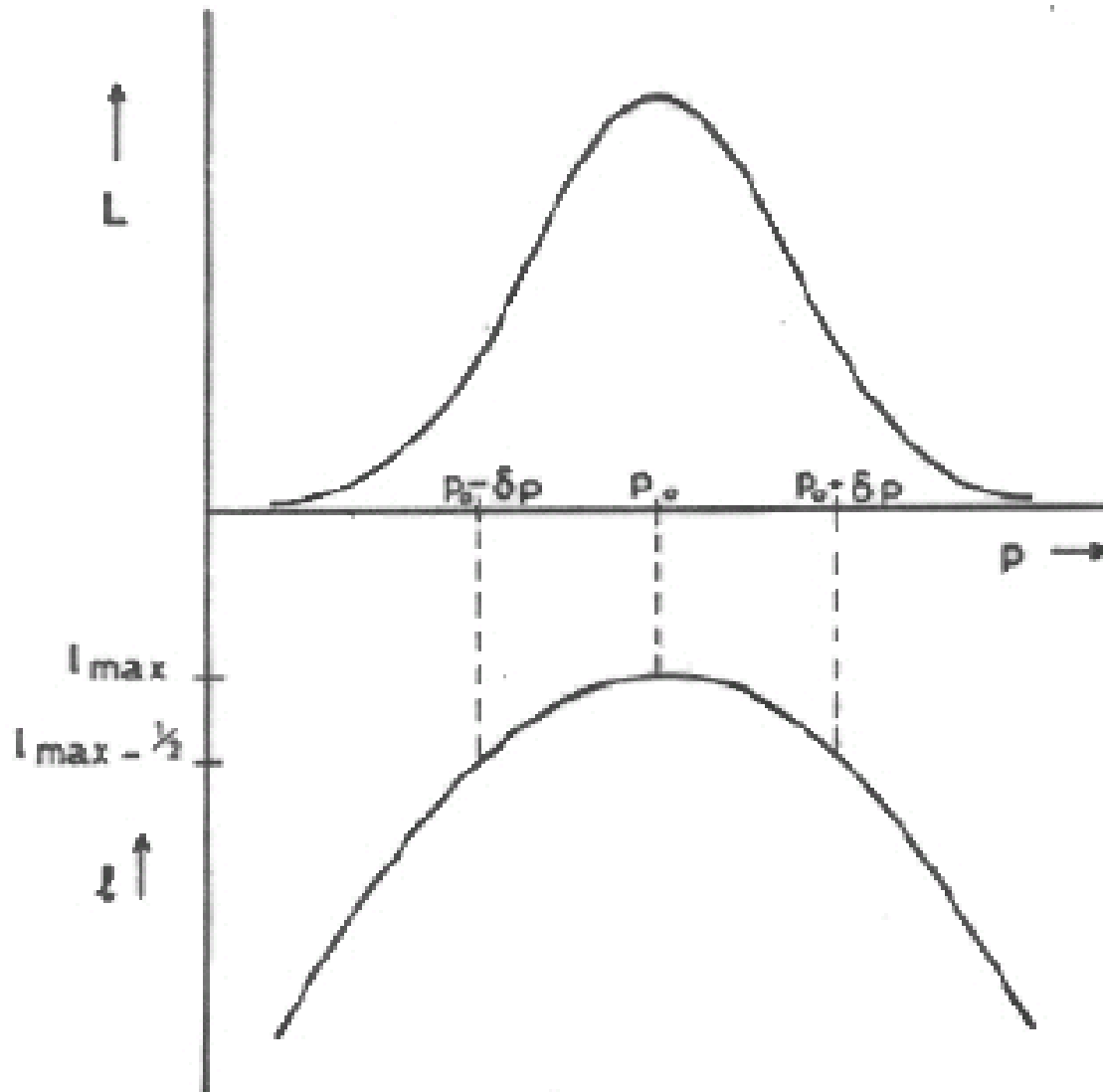
Taylor expand l about its maximum

$$l = l_{\max} + \frac{1}{2!} l'' \left[\delta \left(\frac{c}{a} \right) \right]^2 + \dots$$

$$= l_{\max} - \frac{1}{2c} \delta^2 + \dots \quad c = -1/l''$$

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





Maximum likelihood error

Range of likely values of param μ from width of \mathcal{L} or 1 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”~~

Errors from 3) usually asymmetric, and asym errors are messy.

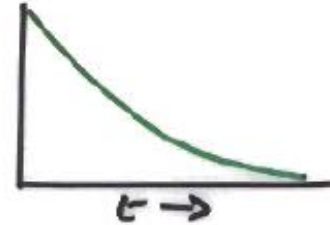
So choose param sensibly

e.g $1/p$ rather than p ; τ or λ

LIFETIME DETERMINATION

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

↑ NORMALISATION



Observe t_1, t_2, \dots, t_N

Use pdf to construct

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

$$\therefore \mathcal{L} = \sum_i (-t_i/\tau - \ln \tau)$$

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

$$\Rightarrow \tau = \sum t_i / N = \bar{t}_i \quad \text{"Obvious"}$$

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

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

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

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

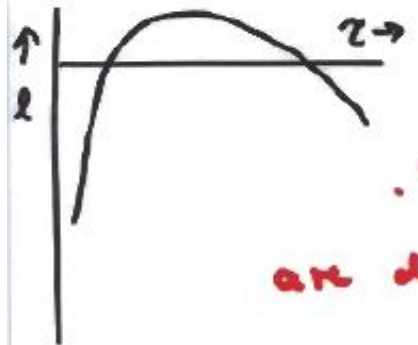
BEWARE FOR AVERAGING RESULTS

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

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

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

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



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

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

— " —

$$l(\tau_{\max}) = -N(1 + \ln \bar{E})$$

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

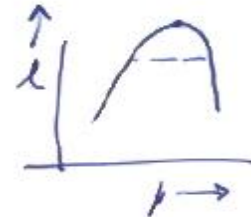
Relevant for whether l_{\max} is useful
for testing goodness of fit

Several Parameters

1 param p

$$p \text{ from } \frac{\partial \ell}{\partial p} = 0$$

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

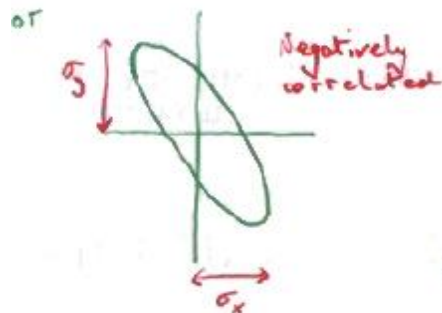
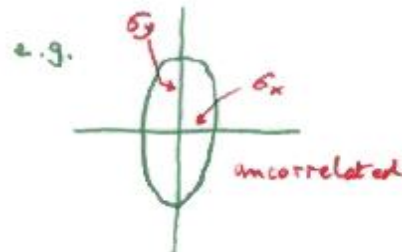


Many dimensions : $\ell(p_1, p_2, p_3, \dots)$

$$p_1, p_2, p_3, \dots \text{ from } \frac{\partial \ell}{\partial p_i} = 0$$

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

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



N.B. ERROR NOT GIVEN BY

$\ell = \ell_{\max} - \frac{1}{2}$ WHEN VARYING x
FROM BEST VALUE WHILE
KEEPING y, \dots CONSTANT

ERROR IS GIVEN BY

$\ell = \ell_{\max} - \frac{1}{2}$ WHEN VARYING x
FROM BEST VALUE WHILE \dots

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

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 $96 \pm 2\%$ protons $4 \pm 2\%$ 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

2) Max Like

Prob for fixed N = Binomial

$$\text{Prob of } f \text{ forwards} \rightarrow f^F (1-f)^B = \frac{N!}{F! B!} \quad *$$

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

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

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

$$\Rightarrow \text{Estimate of } \hat{F} = NF = F \pm \sqrt{FB/N} \leftarrow \text{completely}$$

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

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

$$\text{Maximise } \ln P_b(\nu, f)$$

$$\Rightarrow \hat{\nu} = N \pm \sqrt{N} \leftarrow \text{uncorrelated}$$

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

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 as product of 2 indep Poissons

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

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)

NORMALISATION FOR LIKELIHOOD

$$\int P(x | \mu) dx \quad \text{MUST be independent of } \mu$$

data param

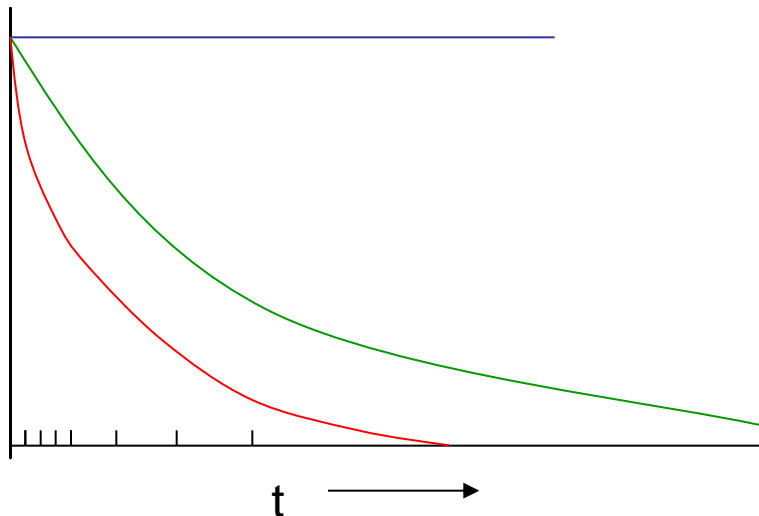
e.g. Lifetime fit to t_1, t_2, \dots, t_n

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

INCORRECT

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

Missing $1/\tau$



— $\tau = \infty$

— τ too big

— Reasonable τ

2) 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

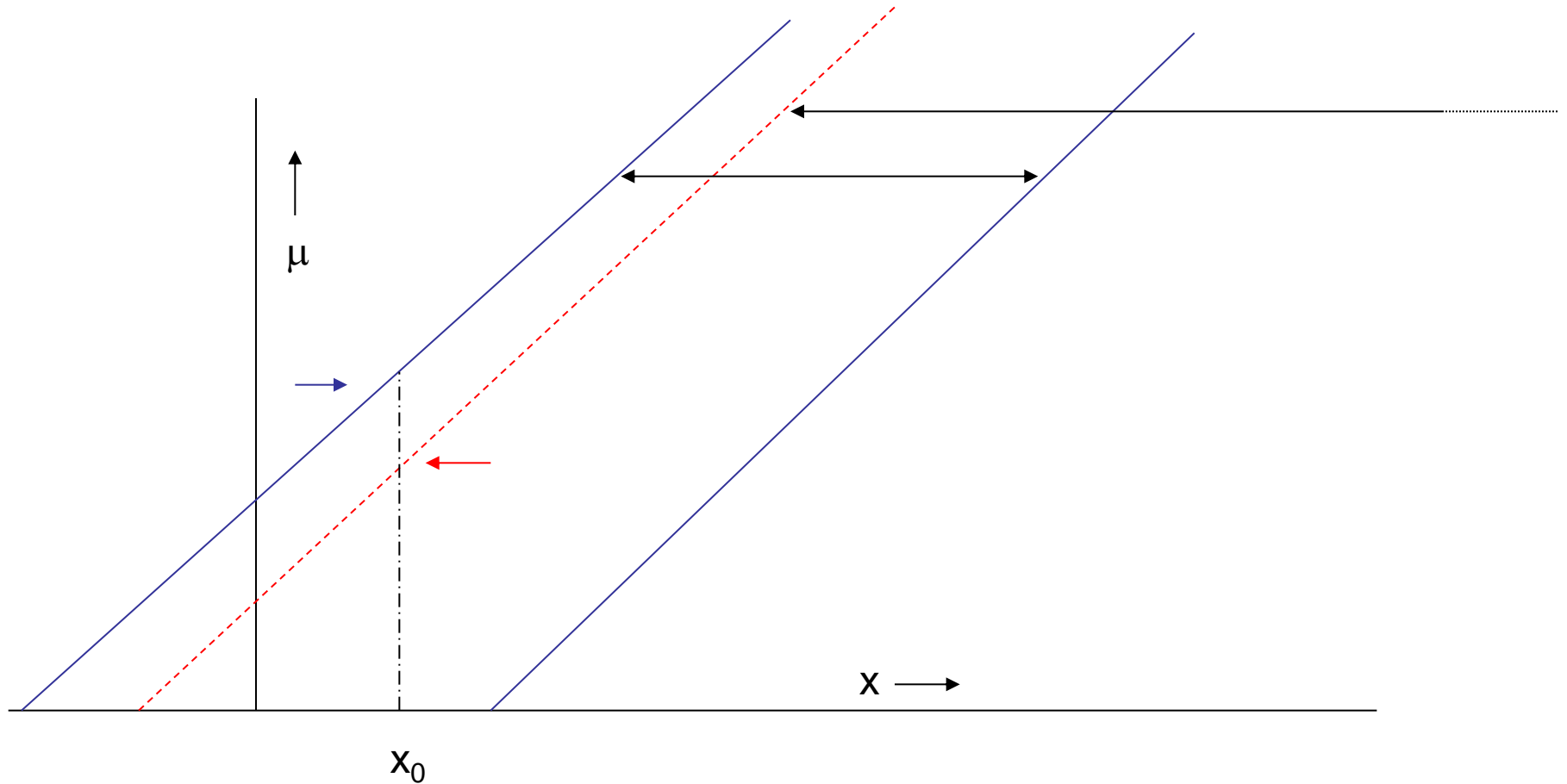
Bayes with prior = const, $1/\mu$ $1/\sqrt{\mu}$ μ etc

“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

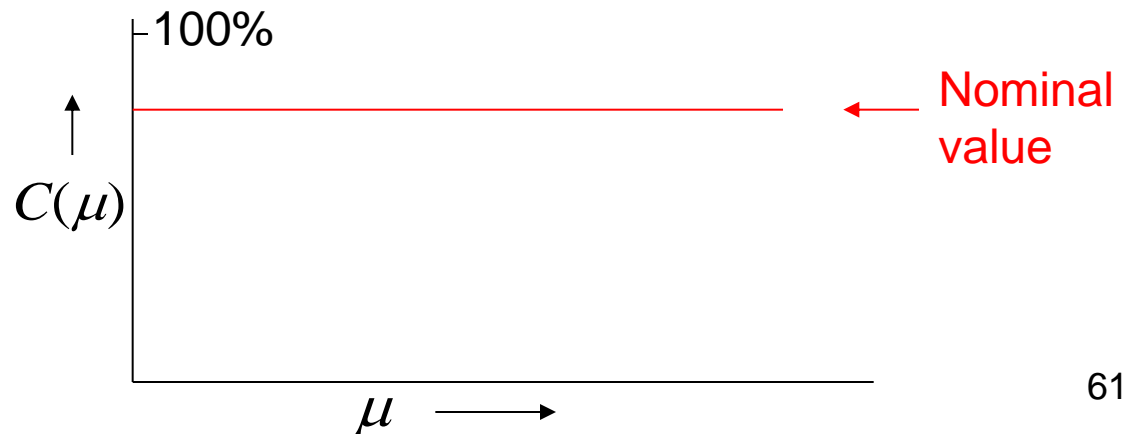
How often does quoted range for parameter include param's true value?

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”

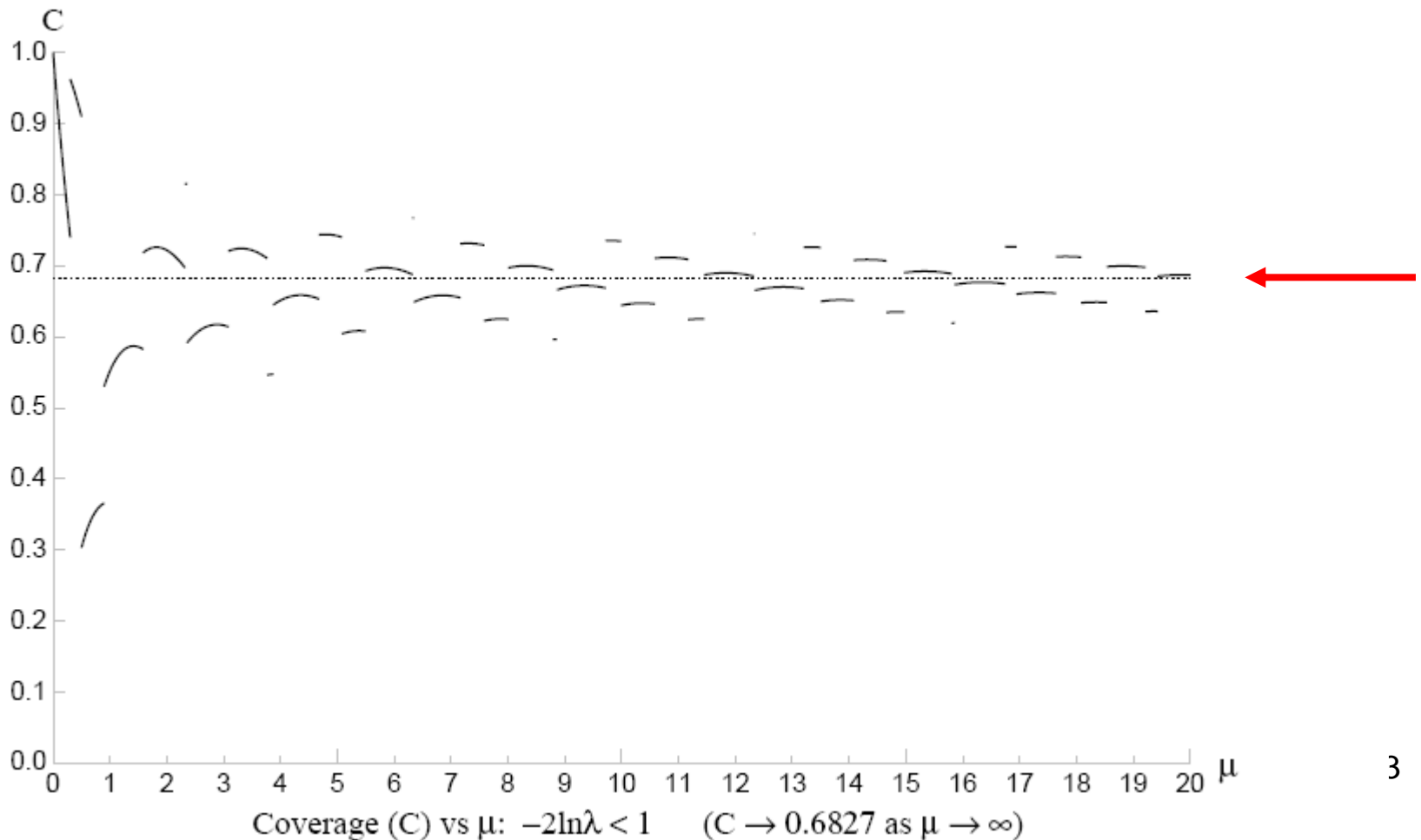
Conservative

Loss of rejection
power

Coverage : \mathcal{L} approach (Not frequentist)

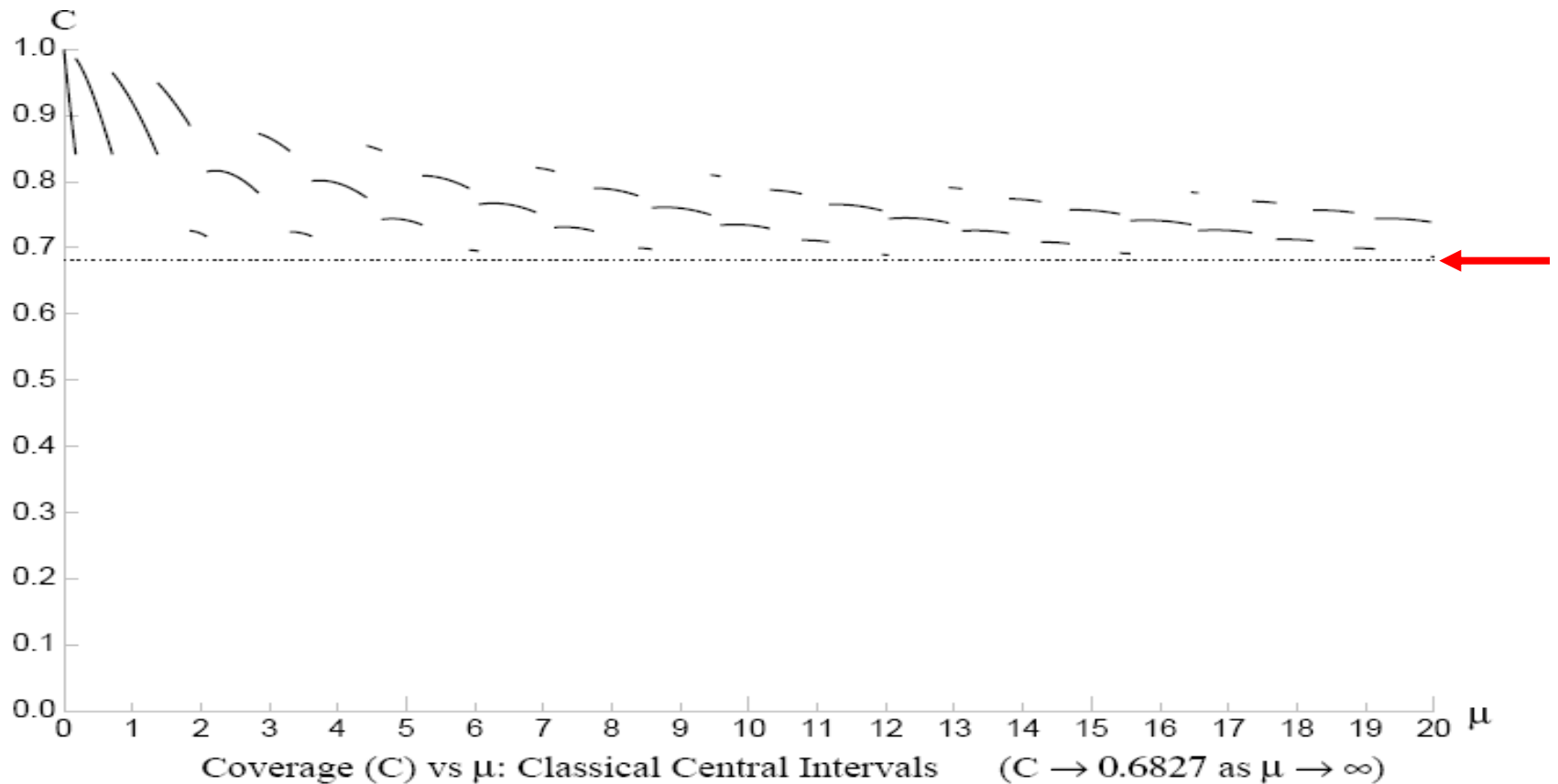
$$P(n, \mu) = e^{-\mu} \mu^n / n! \quad (\text{Joel Heinrich CDF note 6438})$$

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



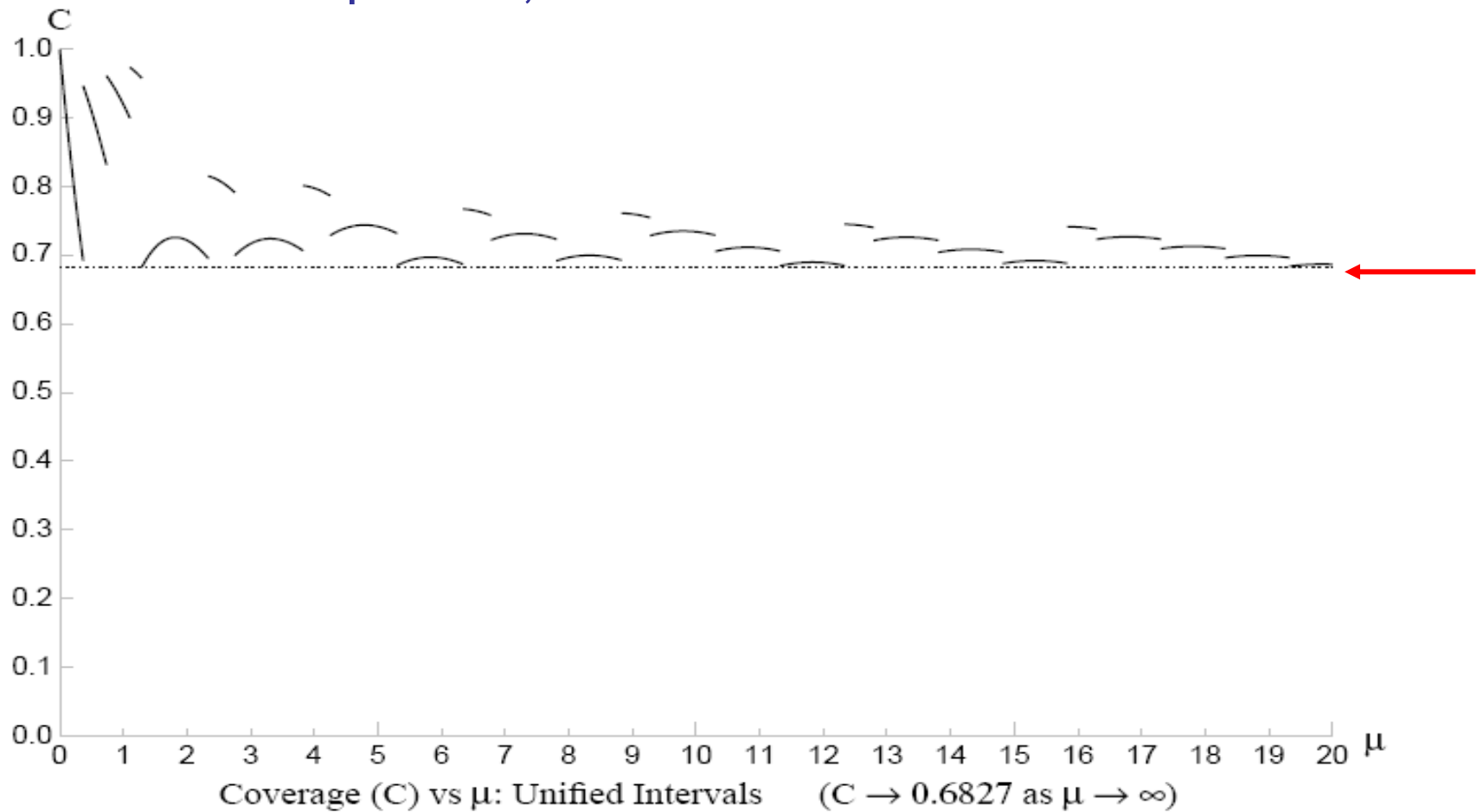
Frequentist central intervals, NEVER undercover

(Conservative at both ends)

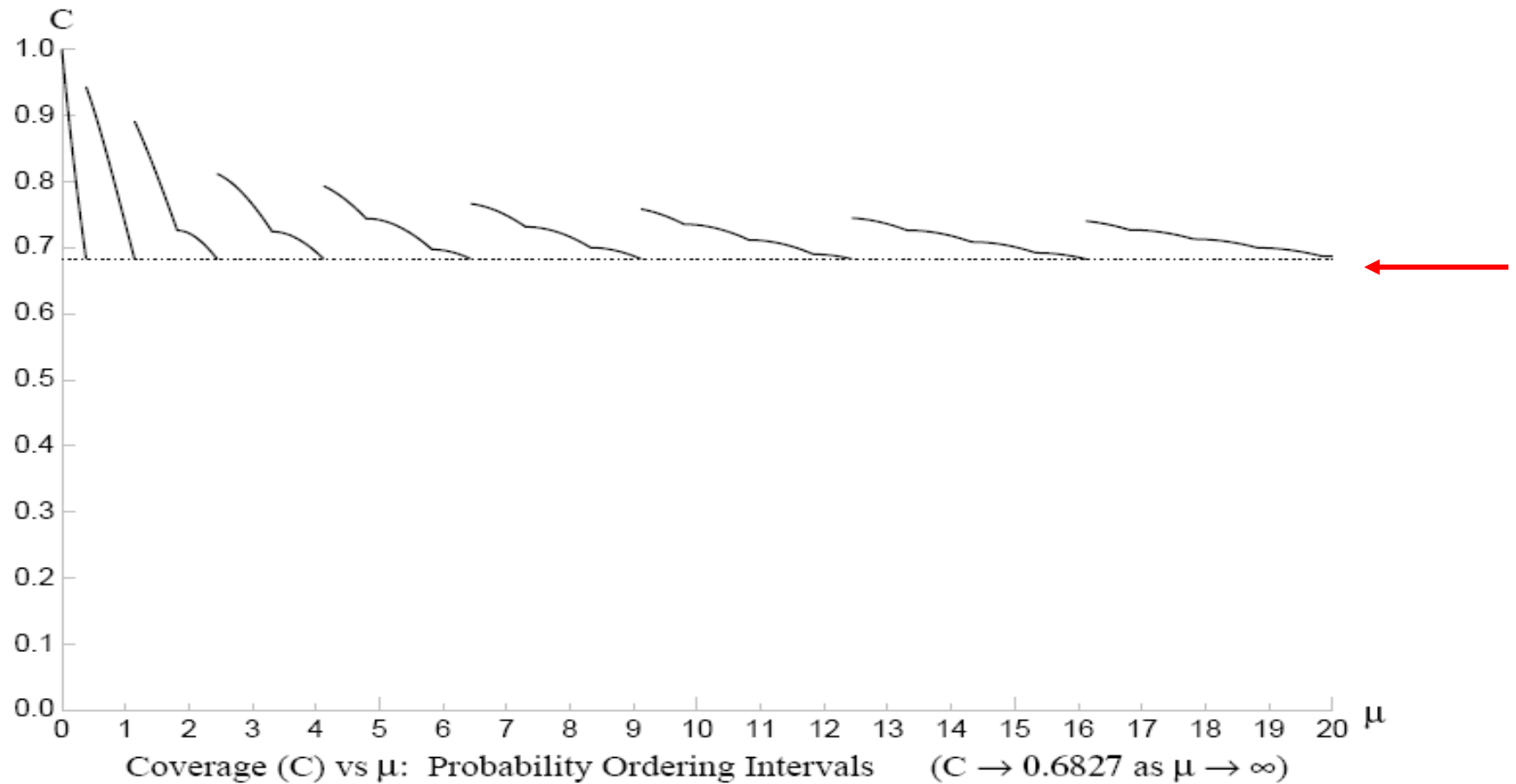


Feldman-Cousins Unified intervals

Frequentist, so NEVER undercovers

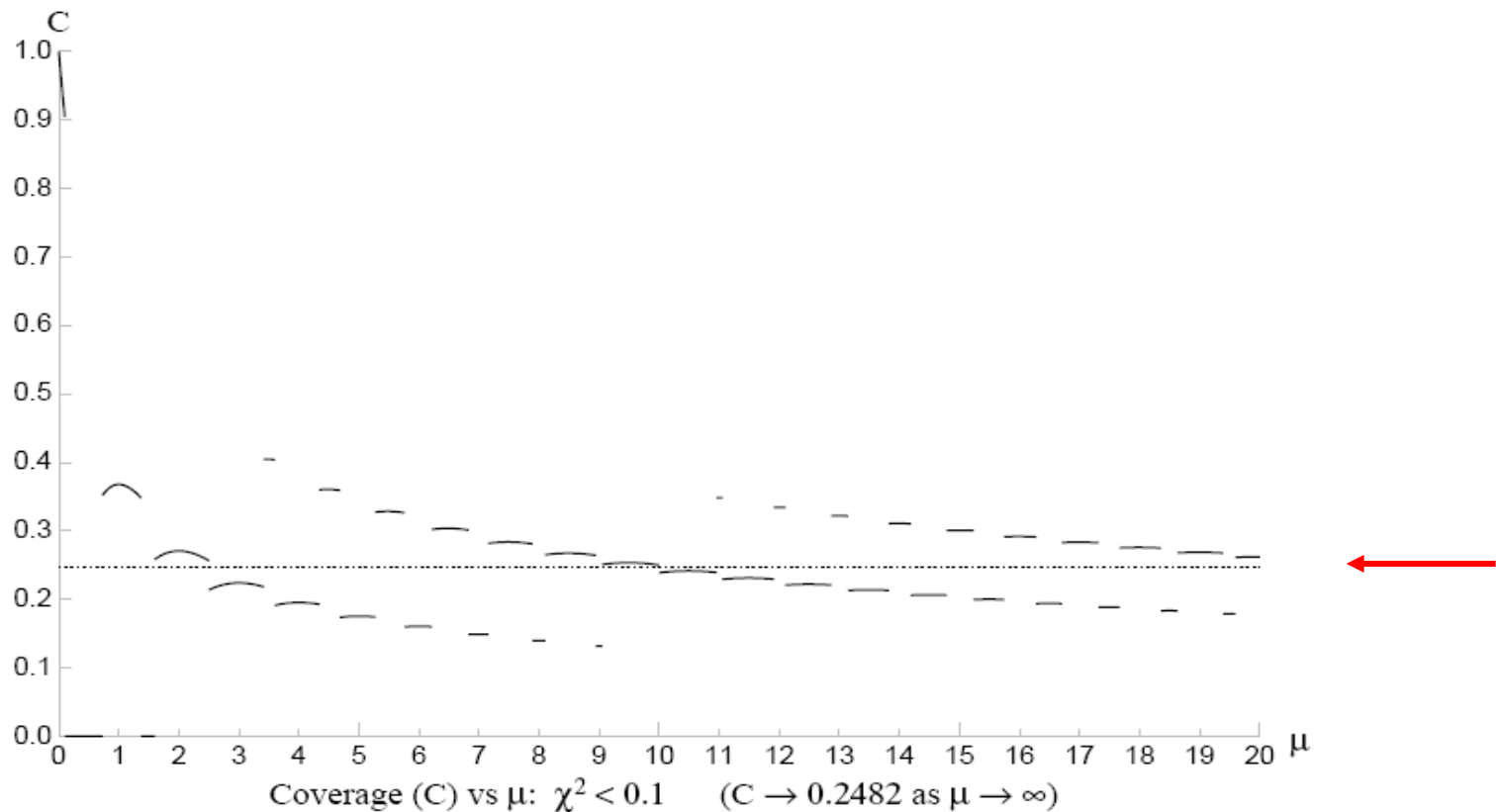


Probability ordering



$$\chi^2 = (n - \mu)^2 / \mu \quad \Delta \chi^2 = 0.1 \quad \longrightarrow \quad 24.8\% \text{ coverage?}$$

NOT frequentist : Coverage = 0% \rightarrow 100%




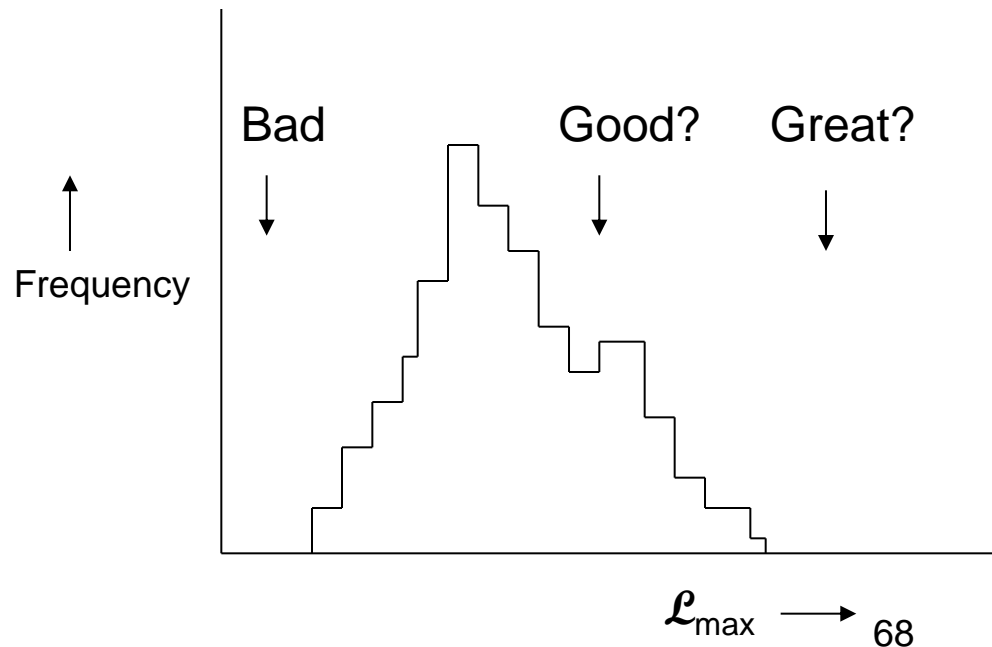
Unbinned \mathcal{L}_{\max} and Goodness of Fit?

Find params by maximising \mathcal{L}

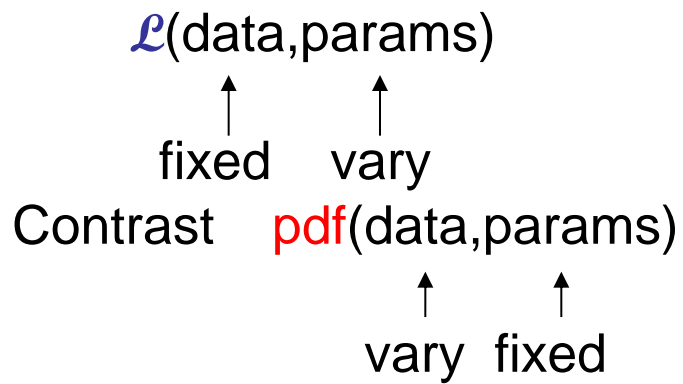
So larger \mathcal{L} better than smaller \mathcal{L}

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

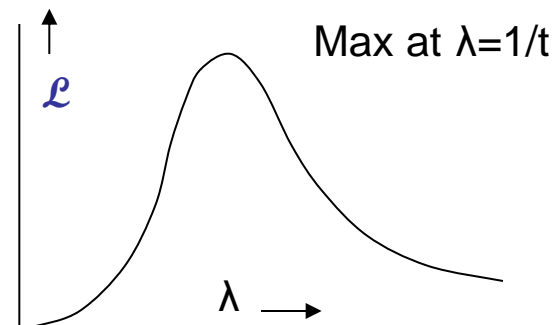
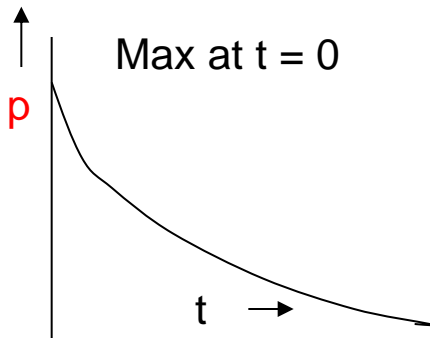
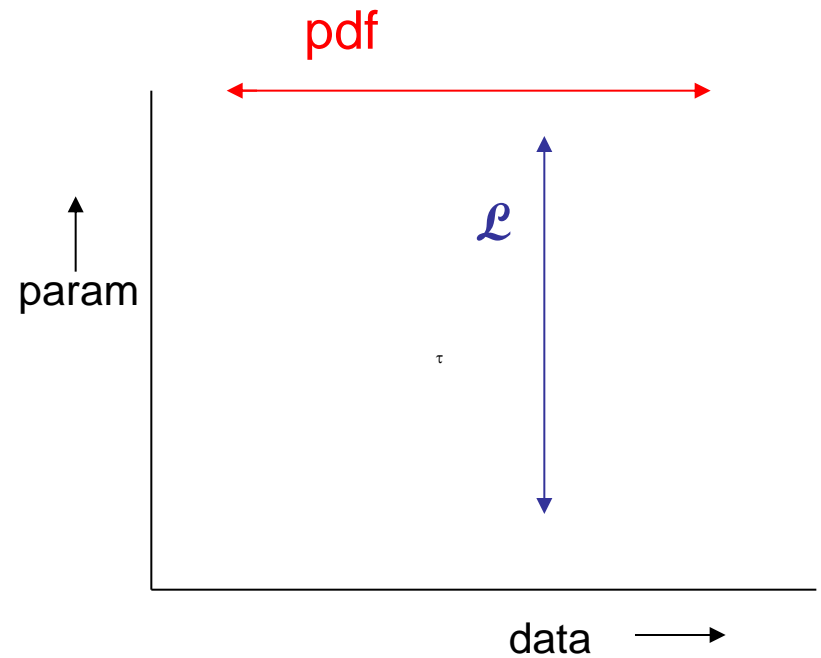
Monte Carlo distribution
of unbinned \mathcal{L}_{\max} 



Not necessarily:



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



Example 1

Fit exponential to times t_1, t_2, t_3, \dots

[Joel Heinrich, CDF 5639]

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

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

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

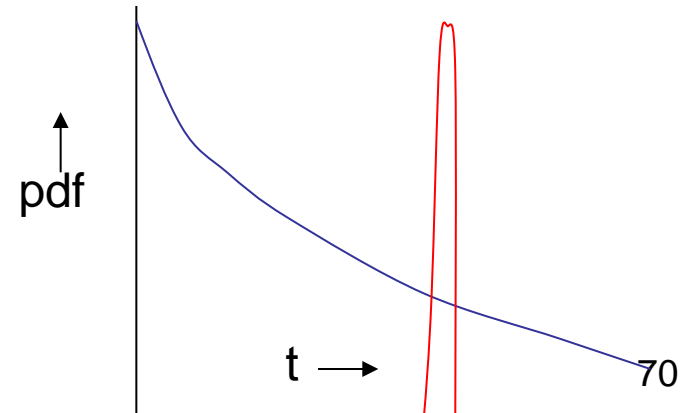
INDEPENDENT OF DISTRIBUTION OF t (except for.....)

(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

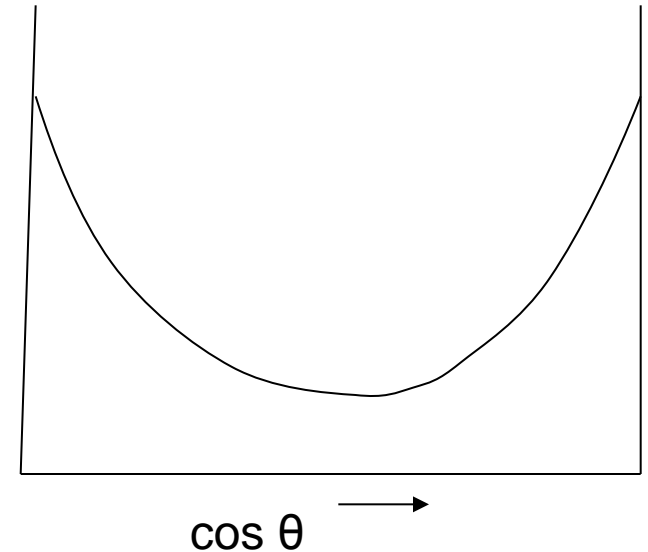
Same average $t \longrightarrow$ same \mathcal{L}_{\max}



Example 2

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

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



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

$$pdf = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\}$$

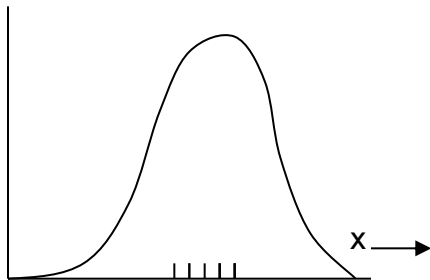
$$\ln \mathcal{L}_{\max} = N(-0.5 \ln 2\pi - \ln \sigma) - 0.5 \sum (x_i - x_{av})^2 / \sigma^2$$

↑ ↑
constant $\sim \text{variance}(x)$

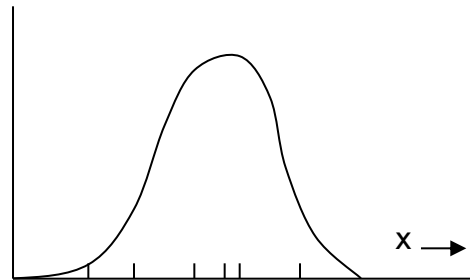
i.e. \mathcal{L}_{\max} depends only on $\text{variance}(x)$,

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

Smaller than expected $\text{variance}(x)$ 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

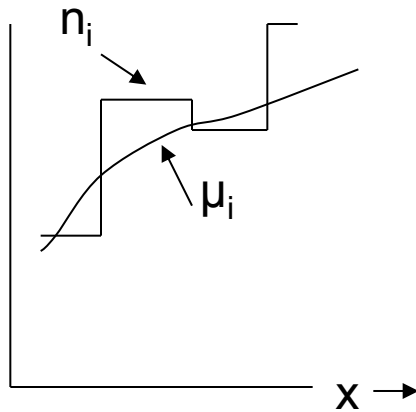
NOT with respect to data

\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



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

$$\begin{aligned}\mathcal{L}_{\text{best}} &= \prod_i P_{n_i}(\mu_{i,\text{best}}) \\ &= \prod_i P_{n_i}(n_i)\end{aligned}$$

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

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

M_{best} is independent of parameters of fit,
and so same parameter values from \mathcal{L} or \mathcal{L} -ratio

\mathcal{L} and pdf

Example 1: Poisson

pdf = Probability density function for observing n , given μ

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

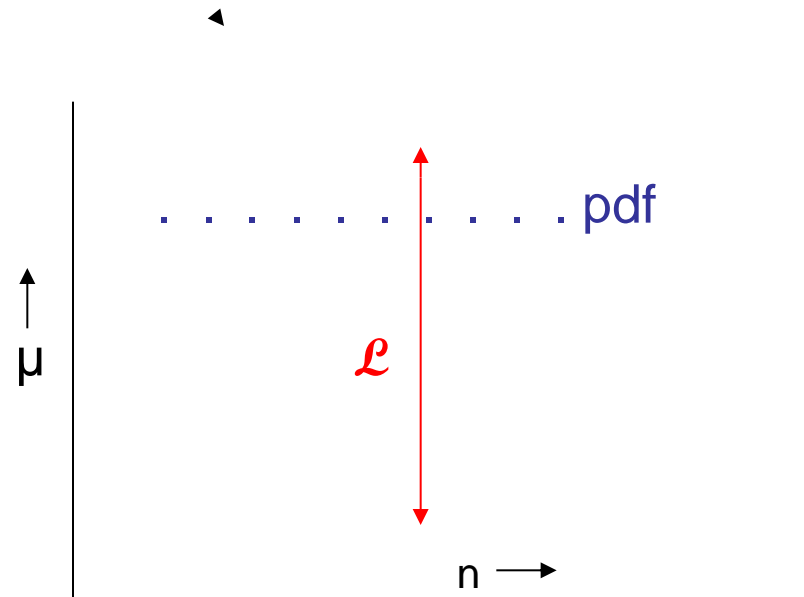
From this, construct \mathcal{L} as

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

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

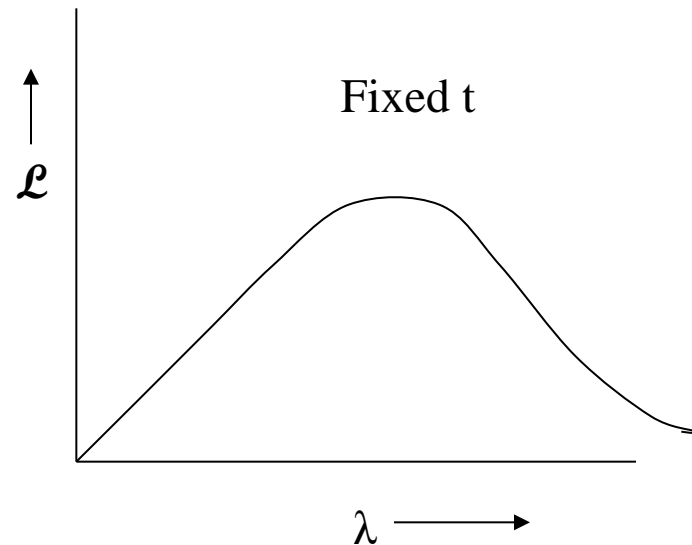
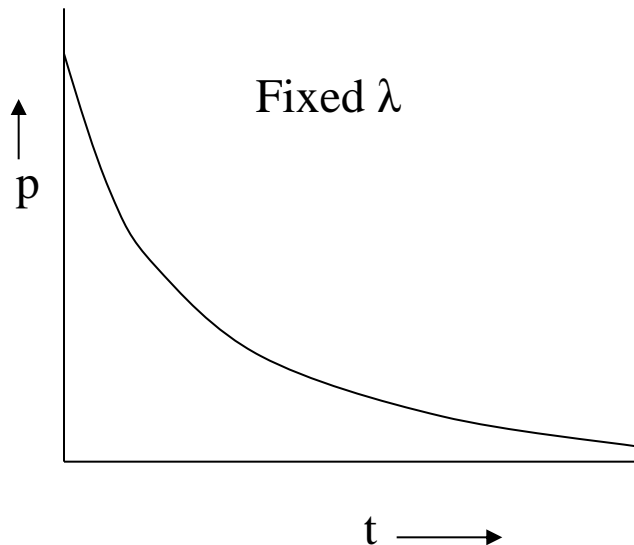
So $L(\lambda;t) = \lambda e^{-\lambda t}$ (single observed t)

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 $L(\lambda)$ are different



Example 3: Gaussian

$$pdf(x; \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

$$L(\mu; x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

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

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

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

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

BUT

$$(b') \quad \int_0^{\lambda_0} L(\lambda;t) d\lambda \neq \int_{\tau_0}^{\infty} L(\tau;t) d\tau$$

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

(However,.....)

	$\text{pdf}(t;\lambda)$	$\mathcal{L}(\lambda;t)$
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:

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

[Metric dependent:

τ range agrees with τ_{pred}

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

BUT

- 1) Could regard as “black box”
- 2) Make respectable by $\mathcal{L} \implies$ Bayes’ posterior

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

6) BAYESIAN SHEARING OF \mathcal{L}

"USE $\ln \mathcal{L}$ FOR $\hat{\mu}$ & σ_p

SHEAR IT TO INCORPORATE
SYSTEMATIC UNCERTAINTIES



SCENARIO:

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

PARAM OF INTEREST

BACKGROUND

EFFICIENCY/ACCEPTANCE/ \mathcal{L}
UNCERTAINTIES
MEASURED IN 'SUBSIDIARY' EXPT

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

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

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

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

i.e. SHEAR \mathcal{L} (not $\ln \mathcal{L}$) 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 \rightarrow different σ_M (or σ_τ)

Events characterised by x_i and σ_i

A events centred on $x = 0$

B events centred on $x = 1$

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

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

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

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

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

So

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

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

but NOT otherwise

Punzi's Monte Carlo for

$$A : G(x, 0, \sigma_A)$$

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

$$f_A = 1/3$$

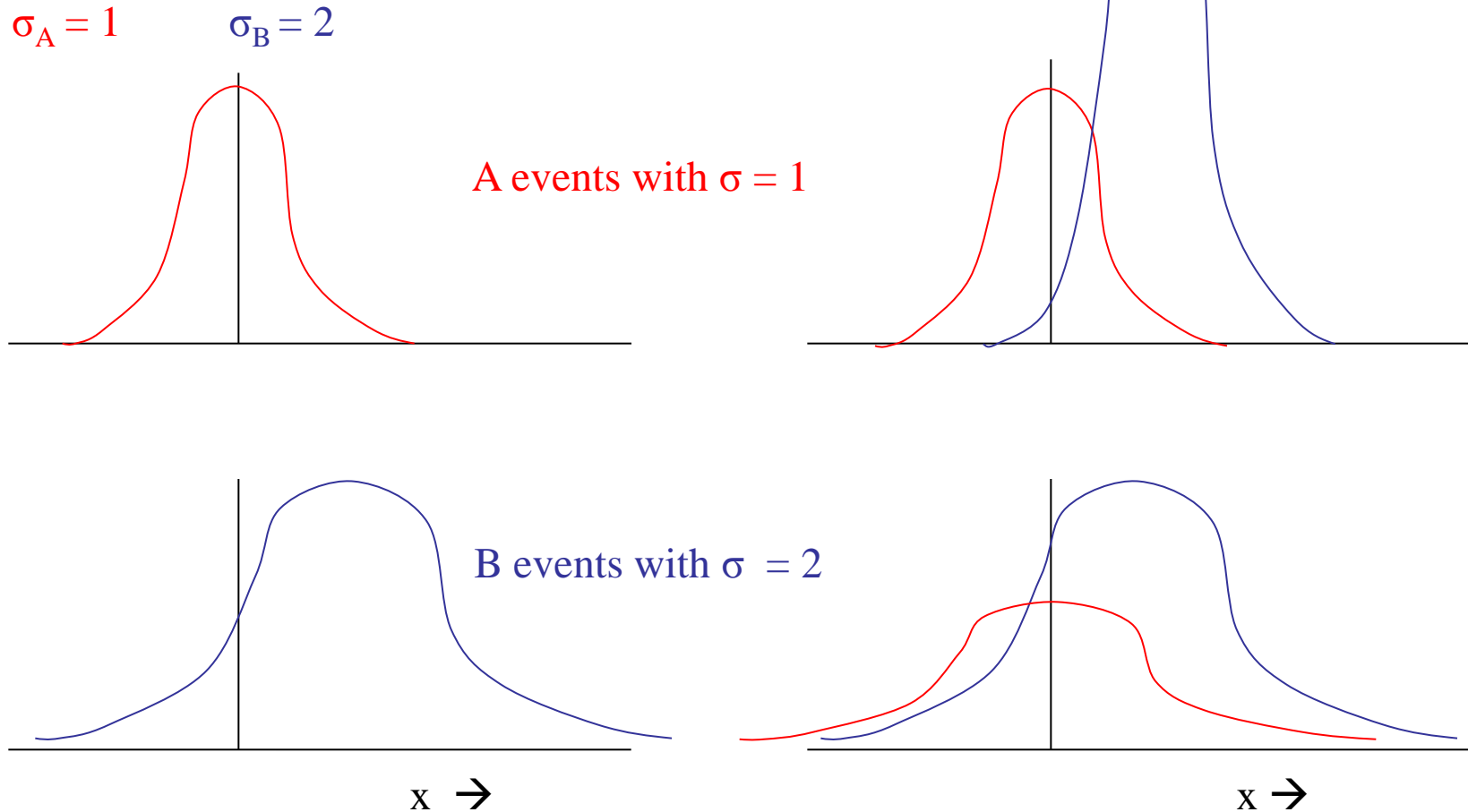
σ_A	σ_B	$\mathcal{L}_{\text{wrong}}$		$\mathcal{L}_{\text{right}}$	
		f_A	σ_f	f_A	σ_f
1.0	1.0	0.336(3)	0.08	Same	
1.0	1.1	0.374(4)	0.08	0.333(0)	0
1.0	2.0	0.645(6)	0.12	0.333(0)	0
1 → 2	1.5 → 3	0.514(7)	0.14	0.335(2)	0.03
1.0	1 → 2	0.482(9)	0.09	0.333(0)	0

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

2) $\mathcal{L}_{\text{right}}$ unbiased, but $\mathcal{L}_{\text{wrong}}$ biased (enormously)!

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

Explanation of Punzi bias



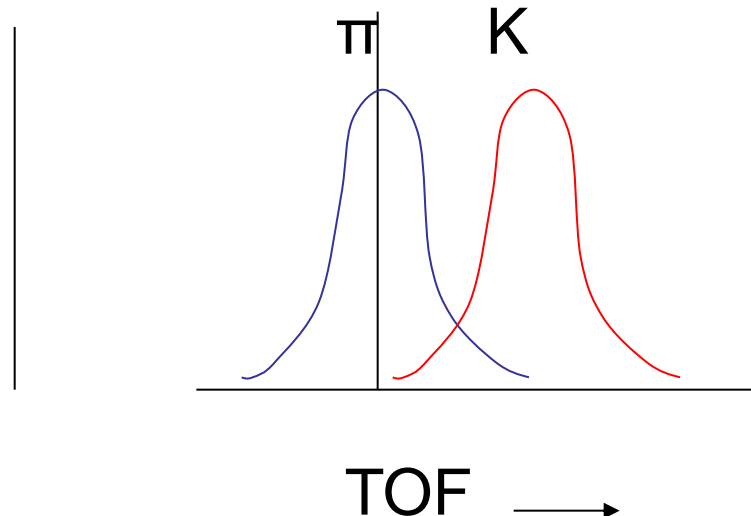
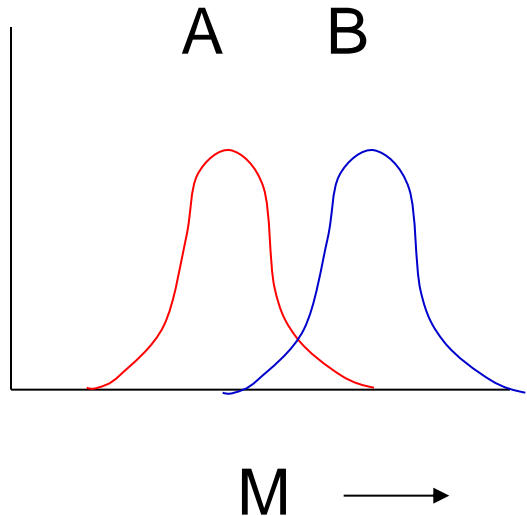
ACTUAL DISTRIBUTION

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

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

K-peak \rightarrow π -peak at large momentum

$\sigma_i \sim \text{constant}$, $p_K \neq p_\pi$

COMMON FEATURE: Separation/Error \neq 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 $(N_A)_i \rightarrow (N_A)_{\text{total}}$, and similarly for B

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

Conclusions

How it works, and how to estimate errors

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

Several Parameters

Likelihood does not guarantee coverage

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

Use correct \mathcal{L} (Punzi effect)

Next time: χ^2 and Goodness of Fit

Least squares best fit

- Resume of straight line

- Correlated errors

- Errors in x and in y

Goodness of fit with χ^2

- Errors of first and second kind

- Kinematic fitting

- Toy example

THE paradox