χ^2 and Goodness of Fit & $\mathcal{L}$ikelihood for Parameters

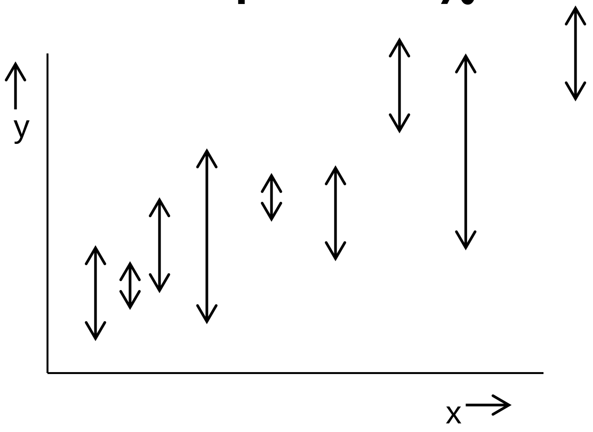
Louis Lyons

Imperial College and Oxford

CERN Summer Students

July 2015

Example of χ^2: Least squares straight line fitting



Data = $\{x_i, y_i \pm \sigma_i\}$

Theory: $y = a + bx$

Statistical issues:

1) Is data consistent with straight line?

(Goodness of Fit)

2) What are the gradient and intercept (and their uncertainties (and correlation))?

(Parameter Determination)

Will deal with issue 2) first

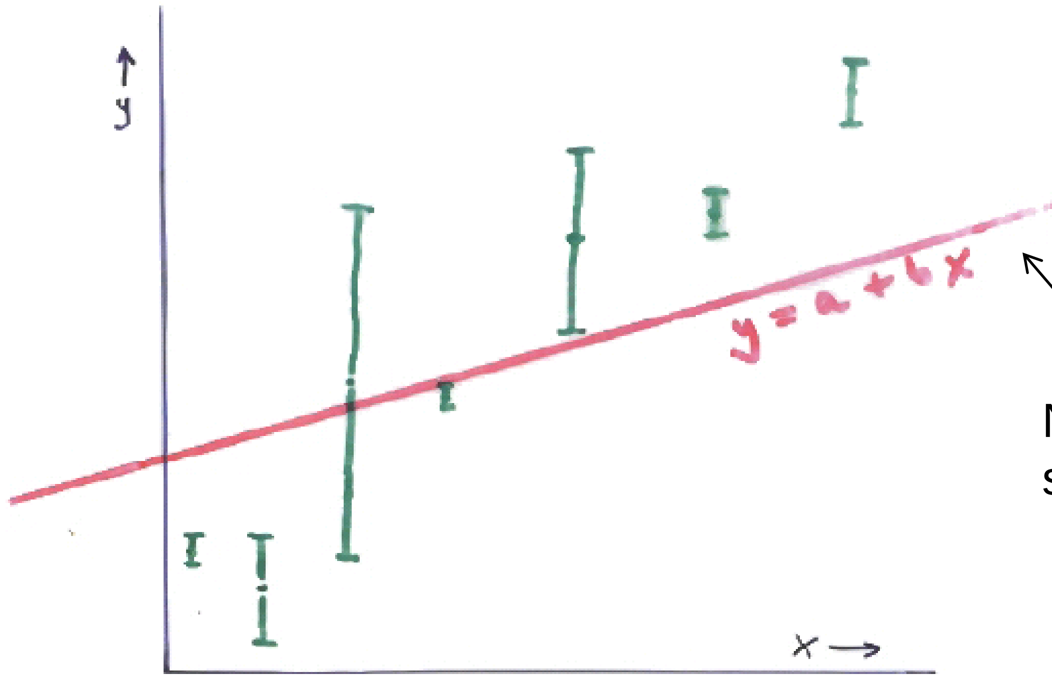
N.B. 1. Method can be used for other functional forms

e.g. $y = a + b/x + c/x^2 + \ldots\ldots$

$y = a + b \sin\theta + c \sin(2\theta) + d \sin(3\theta) + \ldots\ldots$

$y = a \exp(-bx)$

N.B. 2 Least squares is not the only method



Not best
straight line fit

$a + bx_i$

Criterion:

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

An error for each pt.

Minimise S w.r.t. parameters a and b

Straight Line Fit

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

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

ii) Minimise S (w.r.t. $\underline{\underline{a}}$ & $\underline{\underline{b}}$)

$$\left. \begin{aligned} \frac{1}{2}\frac{\partial S}{\partial a} &= \sum_i \frac{(a + b x_i - y_i)}{\sigma_i^2} = 0 \\ \frac{1}{2}\frac{\partial S}{\partial b} &= \sum_i \frac{(a + b x_i - y_i)\, x_i}{\sigma_i^2} = 0 \end{aligned} \right\} \text{2 SIM. EQNS FOR 2 UNKNOWNS } (\underline{\underline{a}} \;\&\; \underline{\underline{b}})$$

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

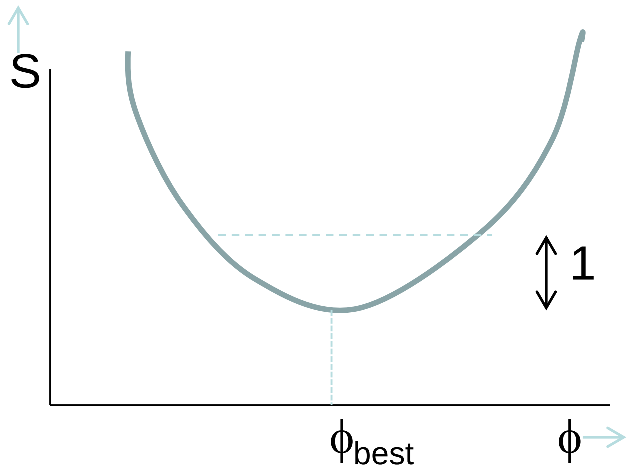
where $[f] = \sum \frac{f_i}{\sigma_i^2}$

& $\langle f\rangle = [f]/[1]$

$\langle y\rangle = a + b\langle x\rangle \Longrightarrow a$

N.B. L.S.B.F. passes through (<x>, <y>)

Uncertainties on parameter(s)

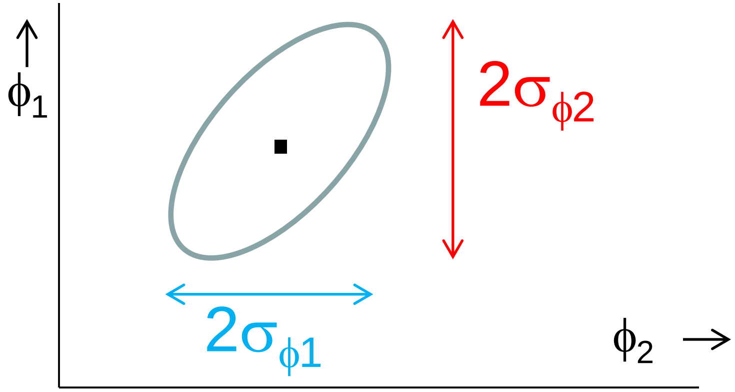


In parabolic approx, $\sigma_\phi = 1/\sqrt{1/2\ d^2S/d\phi^2}$
(mneumonic)

With more than one param, replace $S(\phi)$ by $S(\phi_1, \phi_2, \phi_{3,} \ldots\ldots)$,
and covariance matrix E is given by

$$E^{-1} = \frac{1}{2}\frac{\partial^2 S}{\partial\phi_i \partial\phi_j}$$

$S = S_{max}$ -1 contour



Summary of straight line fitting

- Plot data

  Bad points

  Estimate a and b (and uncertainties)
- a and b from formula
- Errors on a' and b
- Cf calculated values with estimated
- Determine S_{min} (using a and b)
- $\nu = n - p$
- Look up in χ^2 tables
- If probability too small, IGNORE RESULTS
- If probability a "bit" small, scale uncertainties?

Asymptotically

Summary of straight line fitting

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

Parameter Determination

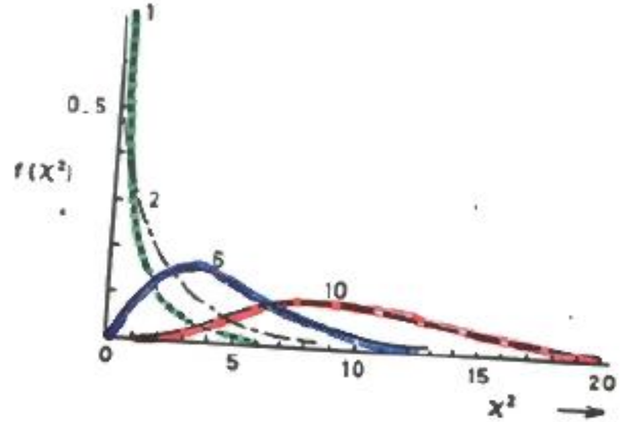
Goodness of Fit

If theory is correct; data unbiassed, ~ Gaussian and asymptotic; σ is correct; etc, then S_{min} has χ^2 distribution

Properties of χ^2 distribution

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

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

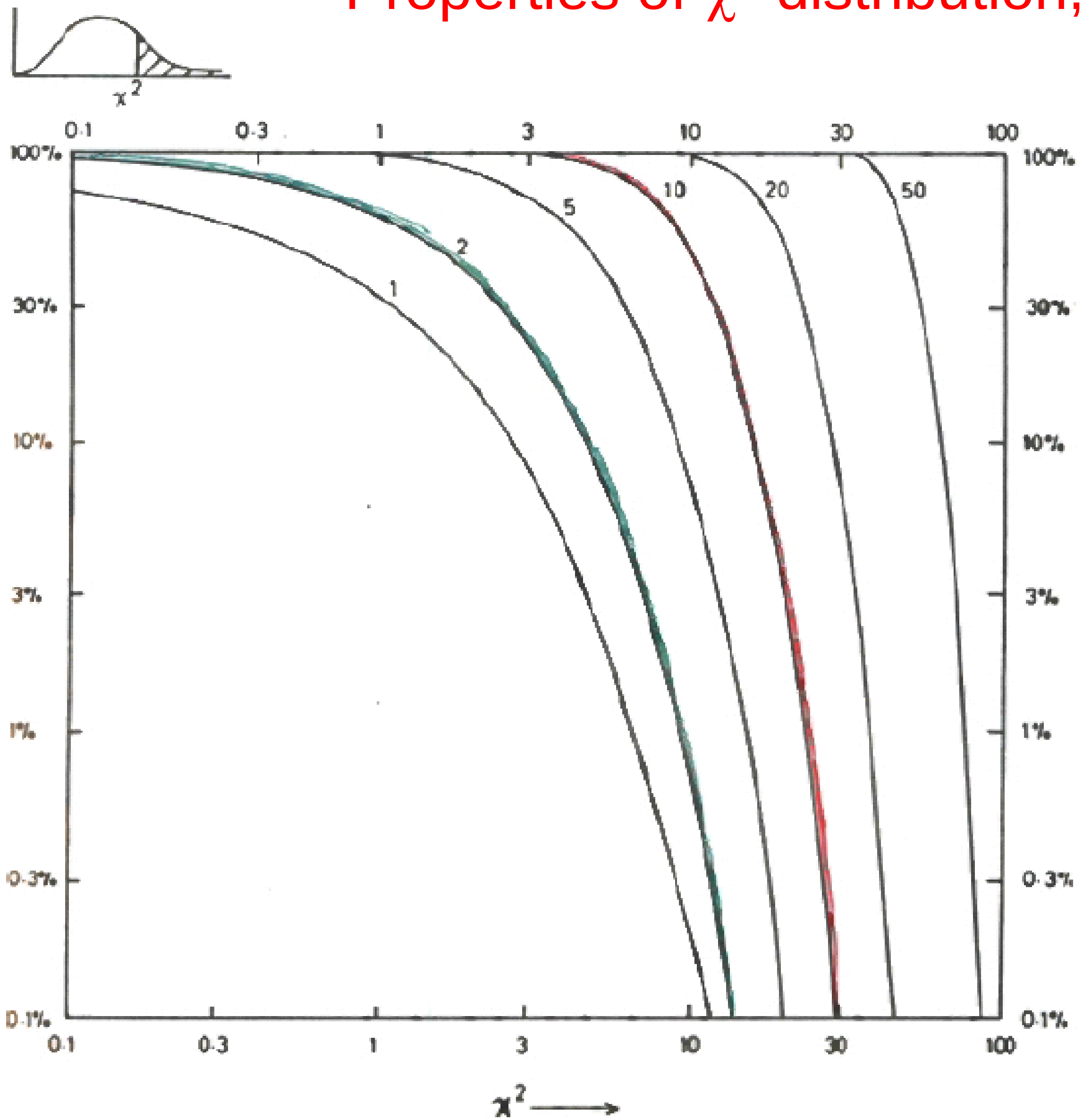


$$\therefore \; S_{min} \gtrsim \nu + 3\sqrt{2\nu}$$

is <u>LARGE</u>

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

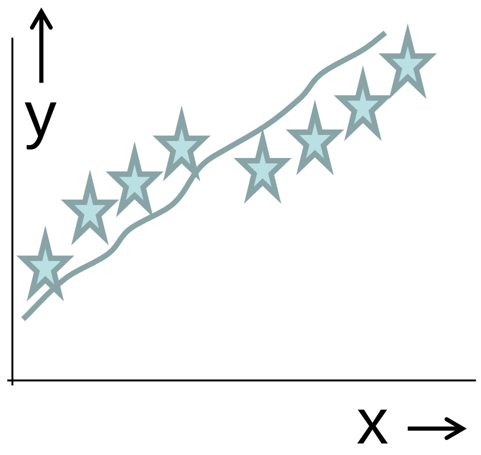
Properties of χ^2 distribution, contd.



CF: Area in tails
of Gaussian

Goodness of Fit

χ^2 Very general
Needs binning
Not sensitive to sign of deviation



Run Test Not sensitive to mag. of devn.

Kolmogorov- Smirnov

Aslan-Zech

Review: Mike Williams, "How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics"
http://arxiv.org/pdf/1006.3019.pdf

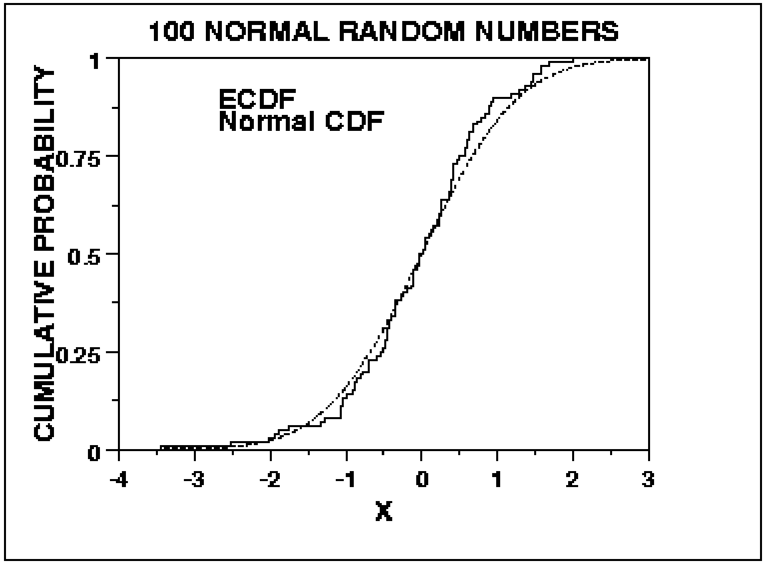
Book: D'Agostino and Stephens, "Goodness of Fit techniques"

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

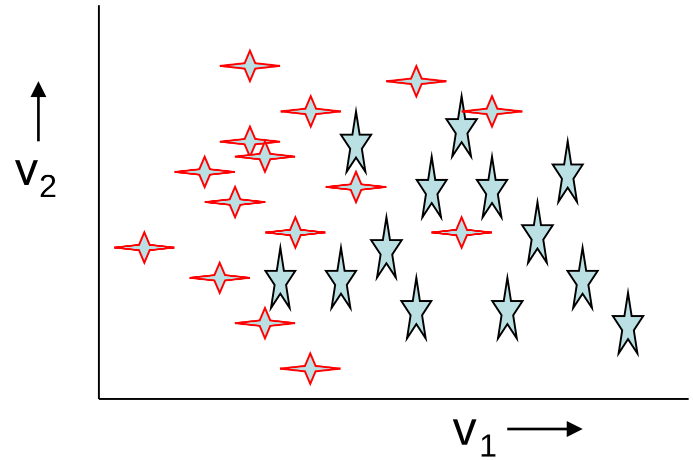
Assign +ve charge to data ; -ve charge to M.C.

Calculate 'electrostatic energy E' of charges

If distributions agree, E ~ 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) $E \sim \Sigma\, q_i q_j\, f(\Delta r = |r_i - r_j|)$, $f = 1/(\Delta r + \varepsilon)$ or $-\ln(\Delta r + \varepsilon)$

   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

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, $S_{min}(p_0) = 90$

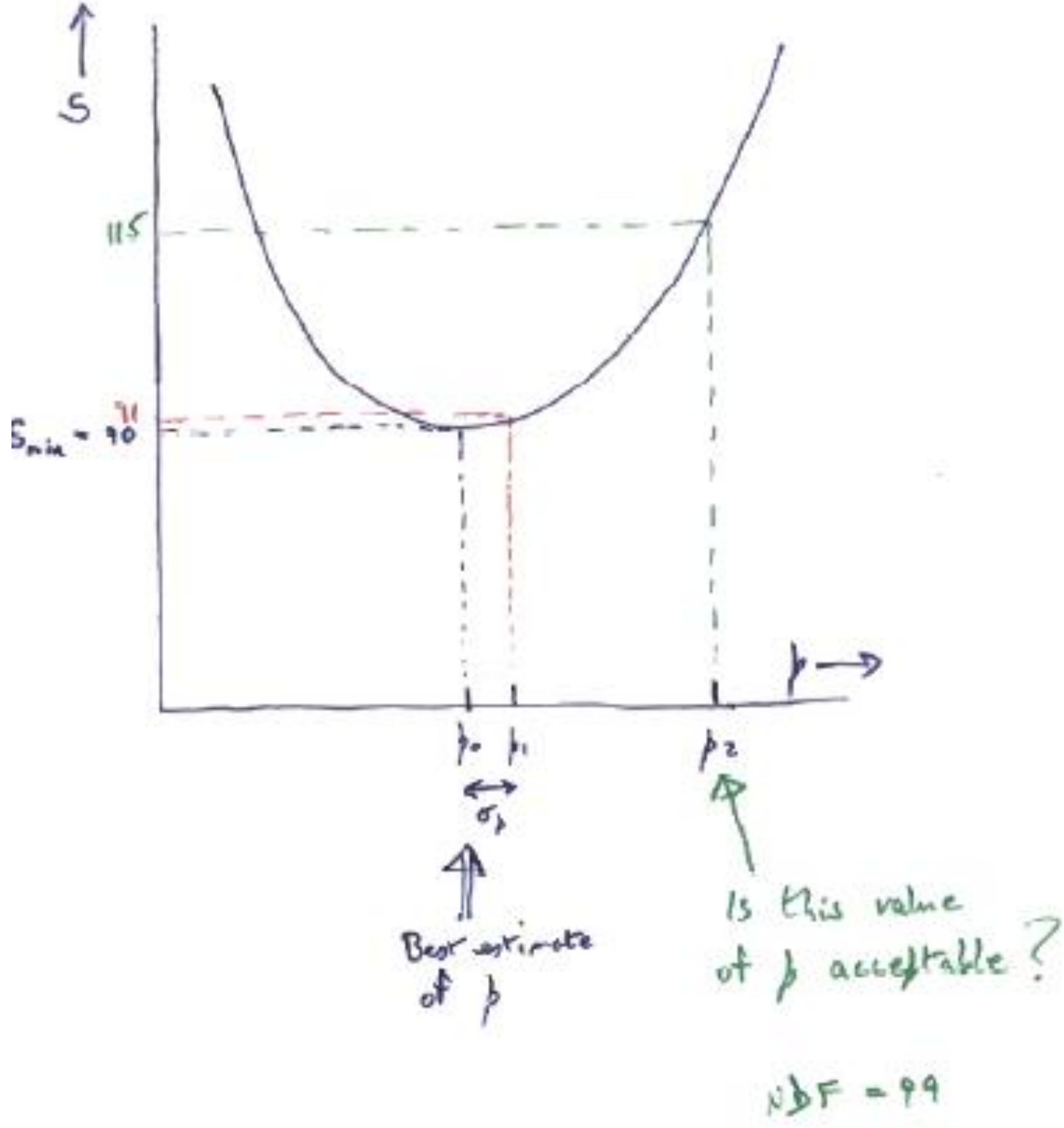
Is p_2 acceptable if $S(p_2) = 115$?

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


S
115
91
S_{min} = 90
p
p_0
p_1
p_2
σ_p
Best estimate
of p
Is this value
of p acceptable?
NDF = 99

$\mathcal{L}$ikelihoods
for determining parameters

What it is

How it works: Resonance

Error estimates

Detailed example: Lifetime

Several Parameters

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

Simple example: Angular distribution

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

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

$\quad$ = probability density of observing θ_i, given β

$L(\beta) = \Pi\, y_i$

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

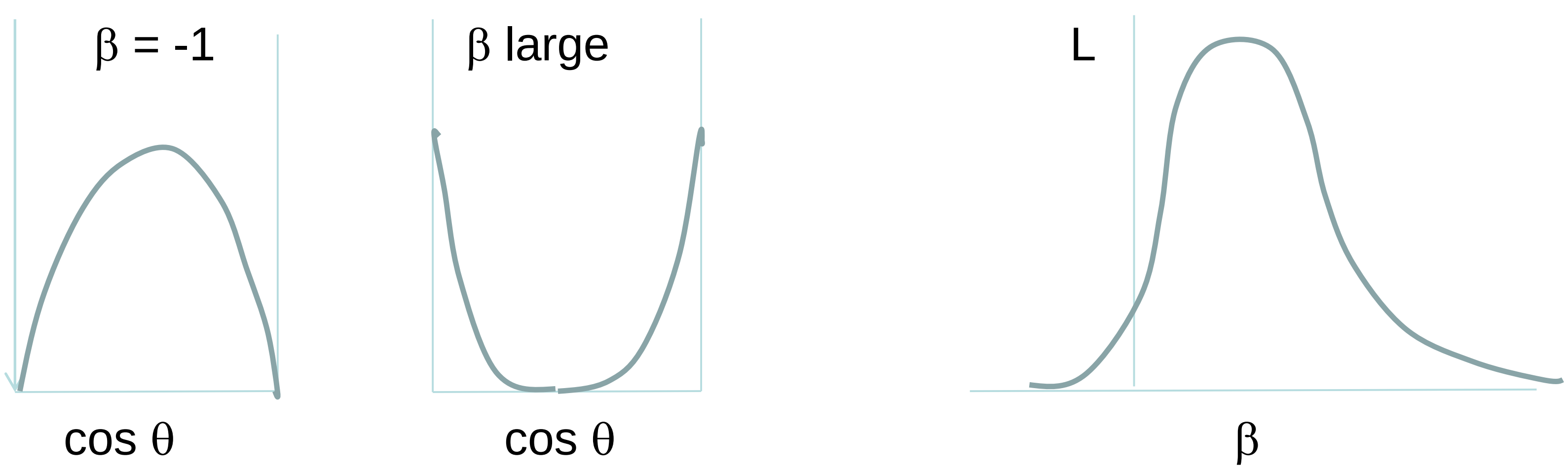
Best estimate of β is that which maximises L

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

Precision of estimate for β comes from width of L distribution

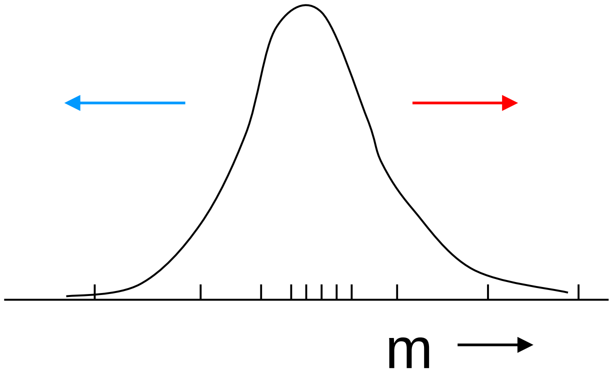
(Information about parameter β comes from shape of exptl distribution of $\cos\theta$)

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

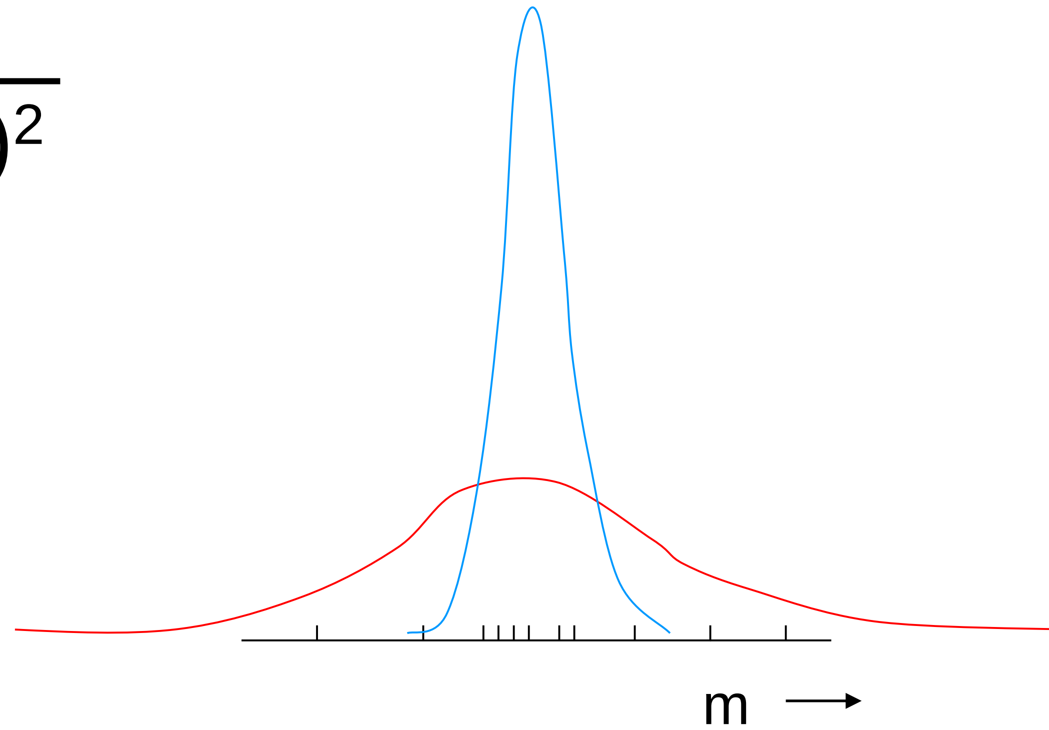


How it works: Resonance

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



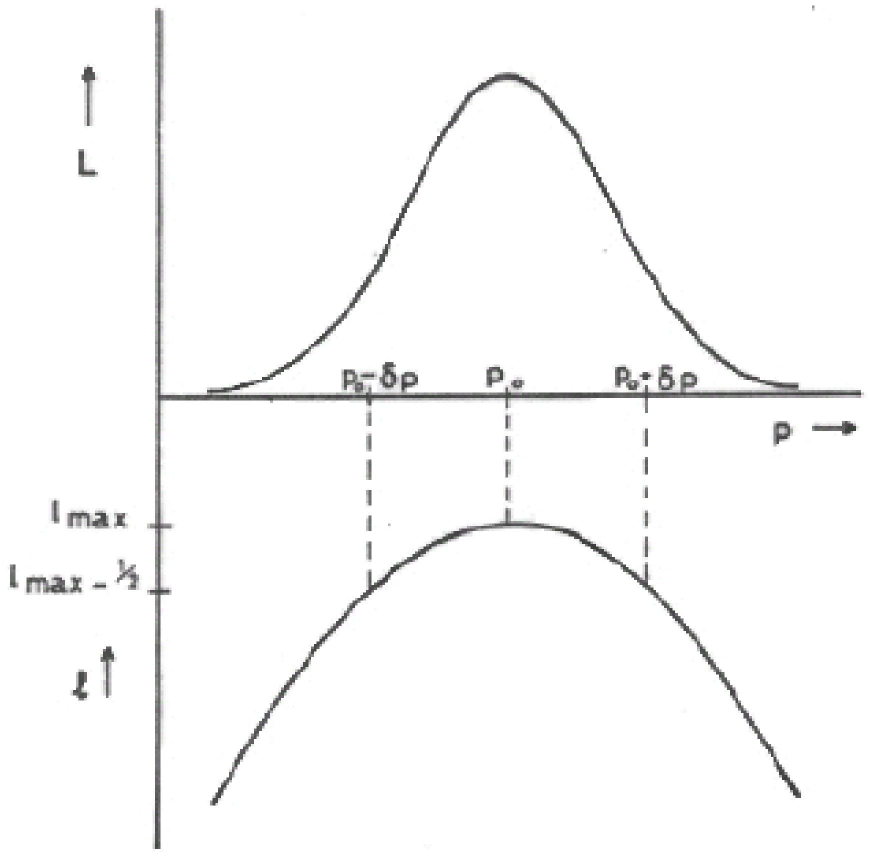
Vary M_0



Vary Γ

Conventional to consider $\boldsymbol{\ell} = \ln(\mathcal{L}) = \Sigma \ln(p_i)$

If $\mathcal{L}$ is Gaussian, $\boldsymbol{\ell}$ is parabolic



Parameter uncertainty with likelihoods

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”~~

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

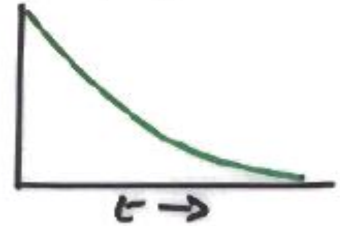
So choose param sensibly

e.g 1/p rather than p; τ or λ

LIFETIME DETERMINATION

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

NORMALISATION



Observe $t_1, t_2 \ldots\ldots t_N$

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"

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

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

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

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

BEWARE FOR AVERAGING RESULTS

	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
Uncertainty estimates	Observed spread, or analytic	$\left\{ \frac{-\partial^2 l}{\partial p_i \partial p_j} \right\}$	$\left\{ \frac{\partial^2 S}{2\partial p_i \partial p_j} \right\}$
Main feature	Easy	Best for params	Goodness of Fit

NORMALISATION FOR LIKELIHOOD

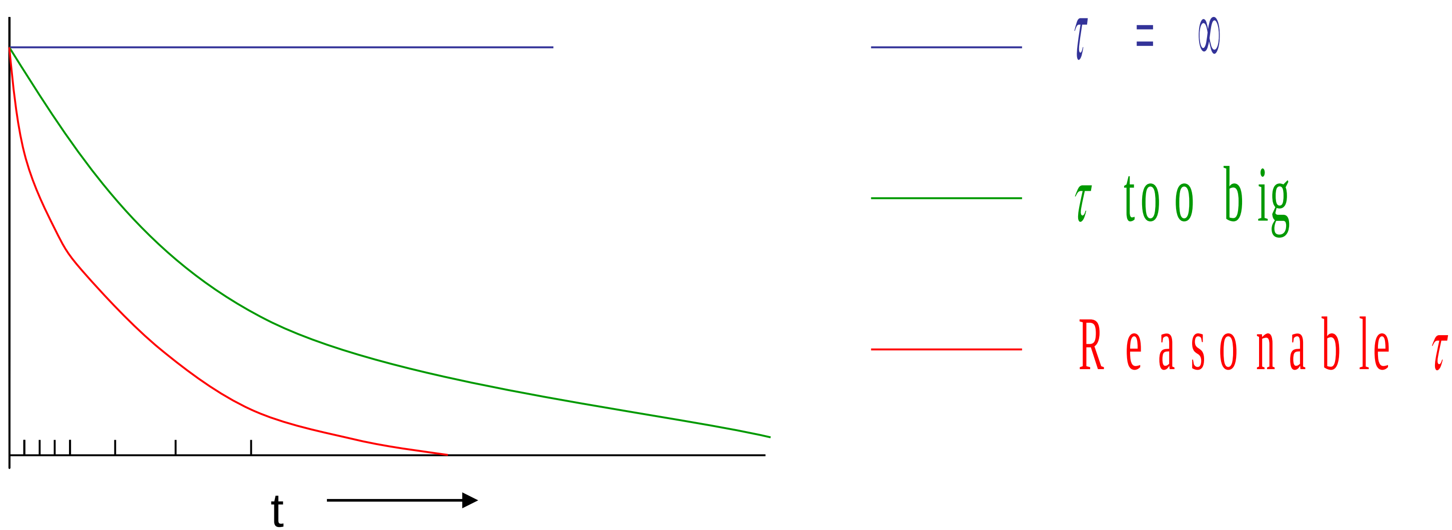
$\int P(x \mid \mu)\, dx$ MUST be independent of μ

data param

e.g. Lifetime fit to $t_1, t_2, \ldots\ldots\ldots t_n$ $[\tau = \sum t_i / N]$

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

Missing $1/\tau$



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

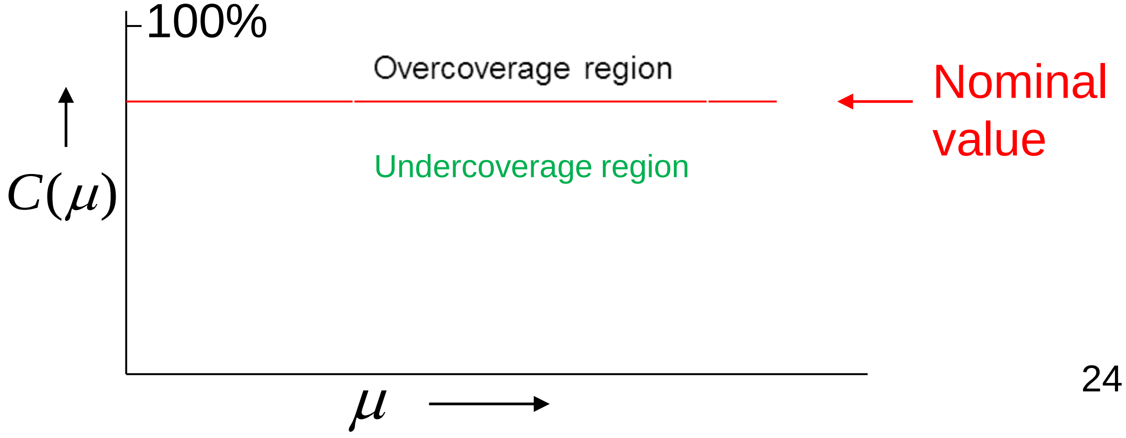
How often does quoted range for parameter include param's true value?

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

Coverage can vary with μ

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

Hope for:



Practical example of Coverage

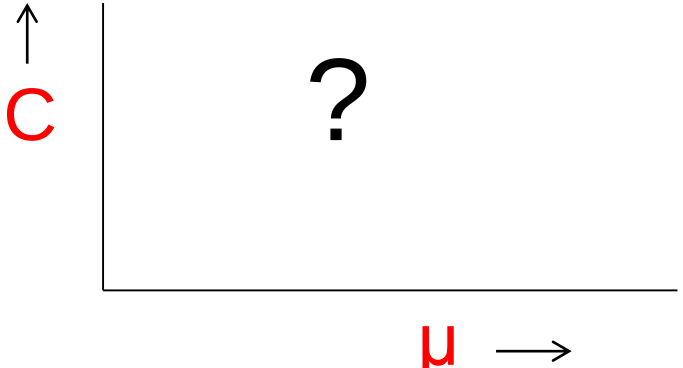
Poisson counting experiment

- Observed number of counts n
- Poisson parameter μ
- $P(n|\mu) = e^{-\mu}\,\mu^n/n!$

Best estimate of μ = n

Range for μ given by $\Delta \ln L = 0.5$ rule. Coverage should be 68%.

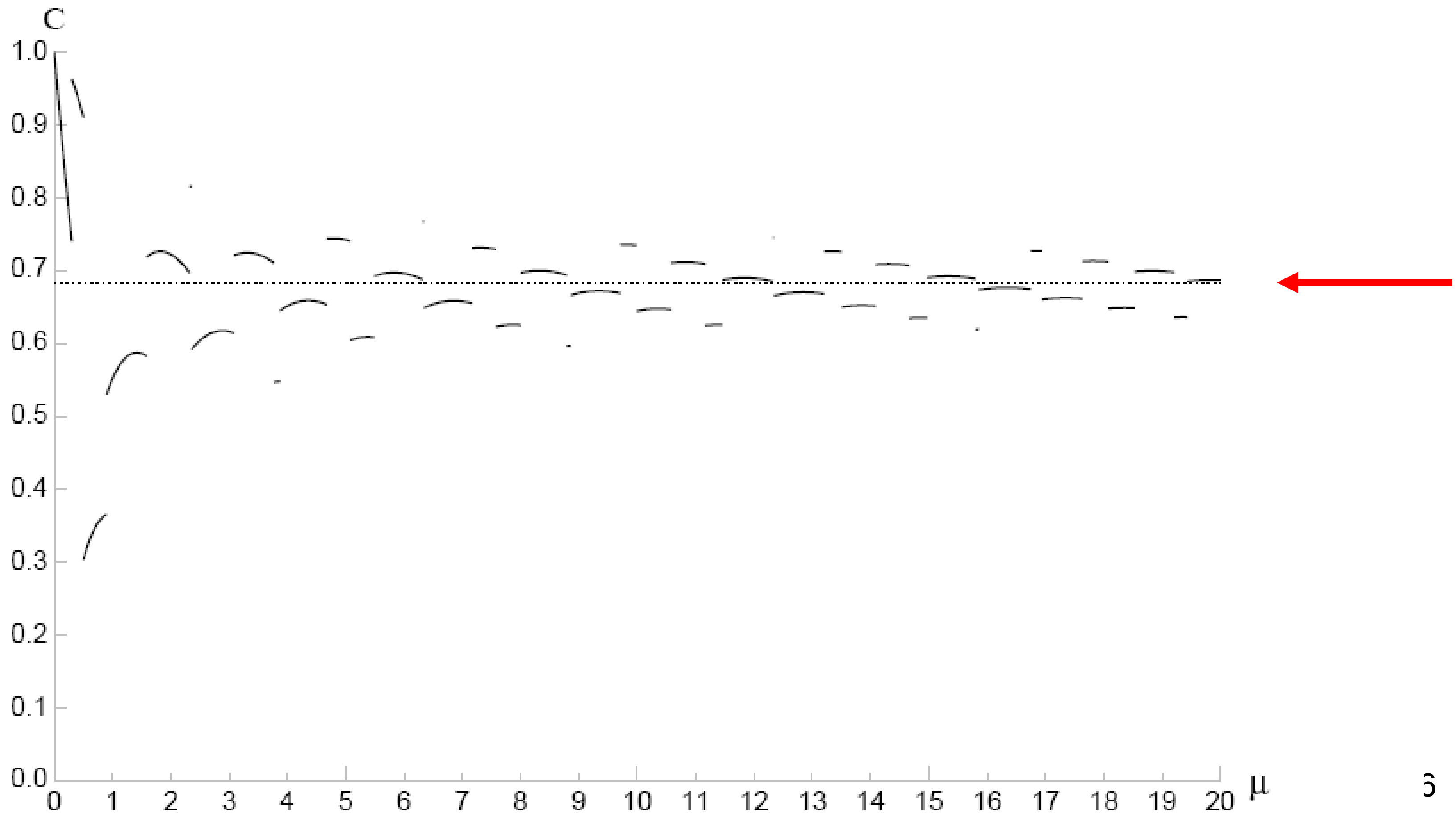
What does Coverage look like as a function of μ?



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

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

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



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

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

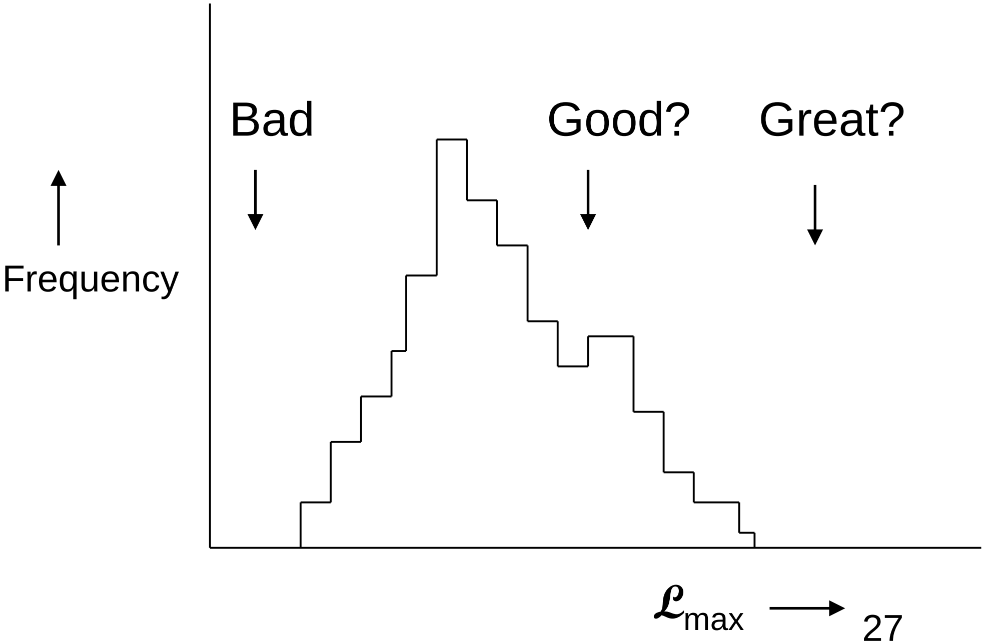
Find params by maximising $\mathcal{L}$

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

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

Monte Carlo distribution

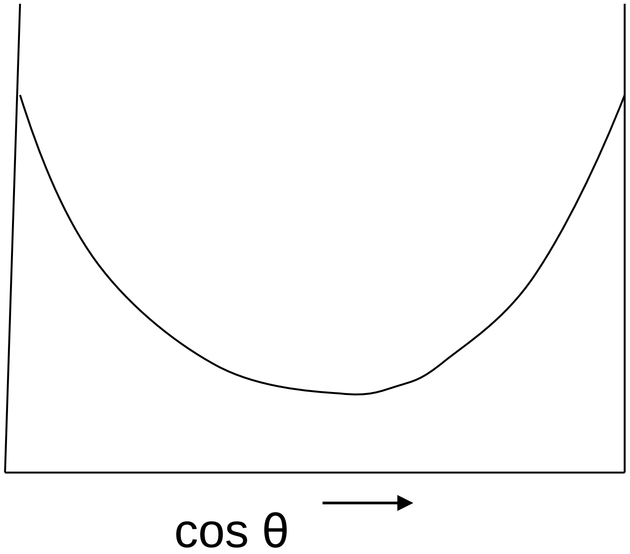
of unbinned $\mathcal{L}_{max}$



Example

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

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



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

Conclusions re Likelihoods

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

Do lifetime and coverage problems on question sheet

Tomorrow

Bayes and Frequentist approaches:
- What is probability?
- Bayes theorem
- Does it make a difference
- Examples from Physics and from everyday life
- Relative merits
- Which do we use?

Last time

Comparing data with 2 hypotheses

H0 = background only (No New Physics)

H1 = background + signal (Exciting New Physics)

Specific example: Discovery of Higgs