PRACTICAL STATISTICS FOR PHYSICISTS

LOUIS LYONS

CDF, OXFORD

LECTURES AT FERMILAB

August 2004

FNAL LECTURE #1

INTRODUCTION

LEARNING TO LOVE THE ERROR MATRIX

STATISTICS FOR NUCLEAR AND PARTICLE PHYSICISTS

L. LYONS

C.U.P (1986)

Available again from C.U.P. ~ 4 July 99

ERRATA in these Lectures

..OTHER BOOKS, ETC

J. OREAR "NOTES ON STATISTICS FOR PHYSICISTS" UCRL-8417 (1958)

D J HUDSON "Lectures on elementary statistics + prob." + "Max like + least squares theory" CERN reports 63-29 + 66-1

S. BRANDT STATISTICAL & COMPUTATIONAL METHODS IN DATA ANALYSIS (North Holland 1973)

W T EADIE et al STATISTICAL METHODS IN EXPTL PHYSICS (North Holland 1971)

S L MEYER DATA ANALYSIS FOR SCIENTISTS + ENGINEERS (Wiley 1975)

A FRODESON et al PROBABILITY + STATISTICS IN PARTICLE PHYSICS (Bergen 1979)

R. BARLOW ~STATISTICS (Wiley, 1993)

G COWAN, STATISTICAL DATA ANALYSIS (Oxford 1998)

B. ROE PROBABILITY & STATISTICS IN EXPTL PHYSICS (Springer-Verlag 1992)

Particle Data Book

CONDITIONAL PROBABILITY

$$\text{PROB}[A+B] = \frac{N(A+B)}{N_{tot}} = \frac{N(A+B)}{N(B)} \cdot \frac{N(B)}{N_{tot}}$$

$$= P(A|B) \times P(B)$$

If A & B are independent, $P(A|B) = P(A)$

$$\Rightarrow P(A+B) = P(A) \times P(B) \quad , \quad A \& B \text{ indep}$$

e.g. $P[\text{Rainy} + \text{Sunday}] = P(\text{rainy}) \times \frac{1}{7}$ INDEP

$P[\text{Rainy} + \text{December}] \neq P(\text{rainy}) \times \frac{1}{12}$ ~~INDEP~~

$P[E_e \text{ large} + E_\nu \text{ large}] \neq P(E_e \text{ large}) \times P(E_\nu \text{ large})$

~~INDEP~~

$P[\text{Beam part 1 interacts} + \text{Beam part 2 interacts}]$

$= [P(\text{beam particle interacts})]^2$ INDEP

$$\text{Prob}[A+B] = \text{Prob}[A|B] \times \text{Prob}[B]$$

$$= \text{Prob}[B|A] \times \text{Prob}[A]$$

PROBABILITY

STATISTICS

Example : Dice

Given $P(5) = 1/6$,
what is P(20 5's out of 100 trials)?

If unbiassed, what is
P(n evens out of 100 trials)?

Given 20 5's out of 100, what is P(5)? And its error?

Observe 65 evens in 100 trials
Is it unbiassed?
Or is $P(\text{even}) = \frac{2}{3}$?

PROBABILITY STATISTICS

Example : Dice

Given $P(5) = 1/6$.
what is P(20 5's out of 100 trials)?

If unbiassed, what is P(n evens out of 100 trials)?

Given 20 5's out of 100, what is P(5)? And its error?

PARAM DETERM.

Observe 65 evens in 100 trials

Is it unbiassed?

GOODNESS OF FIT

Or is $P(\text{even}) = \frac{1}{2}$?

HYPOTHESIS TESTING

THEORY ⇒ DATA DATA ⇒ THEORY

N.B. PARAM DETERMINATION not sensible if GOODNESS OF FIT is poor/bad

ESTIMATE OF VARIANCE

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

UNBIASSED ESTIMATE OF σ^2

$$= \frac{N}{N-1} \left(\overline{x^2} - \bar{x}^2 \right)$$

USEFUL "ON LINE"

BUT can have numerical problems

For Gaussian x_i:

$$\text{error on } s = \frac{\sigma}{\sqrt{2(N-1)}}$$

e.g. $N = 3 \Rightarrow$ 50% error

$N = 51$ for 10% error

COMBINING EXPERIMENTS

(uncorrelated)

$x_i \pm \sigma_i$

$$\hat{x} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$$

$$1/\sigma^2 = \sum 1/\sigma_i^2$$

From $S = \sum (x_i - \hat{x})^2/\sigma_i^2$

← Minimise S

← σ from $S_{min} + 1$

OR Propagate errors from $\hat{x} = \ldots\ldots$

Define $w_i = 1/\sigma_i^2$ = weight ~ information content

$$\hat{x} = \sum w_i x_i / \sum w_i$$

$$W = \sum w_i$$

Example: Equal σ_i $\Rightarrow$ $\hat{x} = \bar{x}$

$\sigma = \sigma_i/\sqrt{n}$

BEWARE

100 ± 10
1 ± 1
} → 2 ± 1 ?

or 50.5 ± 5 ?

DIFFERENCE BETWEEN ADDING + AVERAGING

No of MARRIED MEN = 10.0 ± 0.5 Million

No of MARRIED WOMEN = 8 ± 3 Million

Total = 18 ± 3 million

Average = 9.9 ± 0.5

$\Rightarrow$ Total = 20 ± 1 million

General point: Including theoretical input can improve accuracy of answer

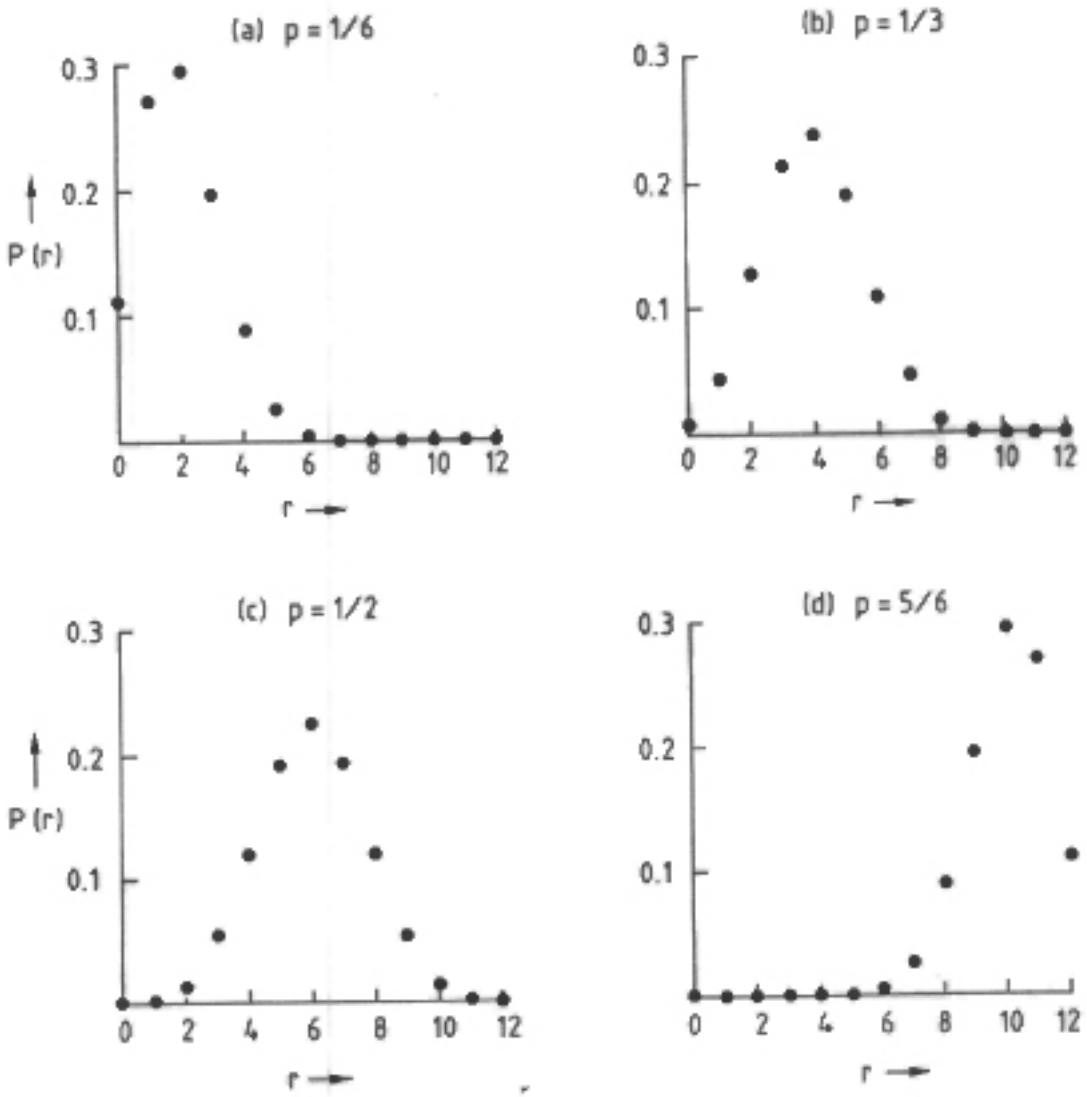


Fig. A3.1 The probabilities $P(r)$, according to the binomial distribution, for r successes out of 12 independent trials, when the probability p of success in an individual trial is as specified in the diagram. As the expected number of successes is $12p$, the peak of the distribution moves to the right as p increases. The RMS width of the distribution is $\sqrt{12p(1-p)}$ and hence is largest for $p = \frac{1}{2}$. Since the chance of success in the $p = \frac{1}{6}$ case is equal to that of failure for $p = \frac{5}{6}$, the diagrams (a) and (d) are mirror images of each other. Similarly the $p = \frac{1}{2}$ situation shown in (c) is symmetric about $r = 6$ successes.

Thus the expected number of successes of our die-throwing experiment was $12 \times (1/6) = 2$, with a variance of $12 \times (1/6) \times (5/6) = 5/3$ (or a standard deviation of $\sqrt{5/3}$). This tells us that we cannot expect that the number of successes will be much larger than a couple of times $\sqrt{5/3}$ above 2, i.e. more than five 6's is unlikely (see Fig. A3.1(a)).

For the same experiment of throwing a die 12 times, we could have

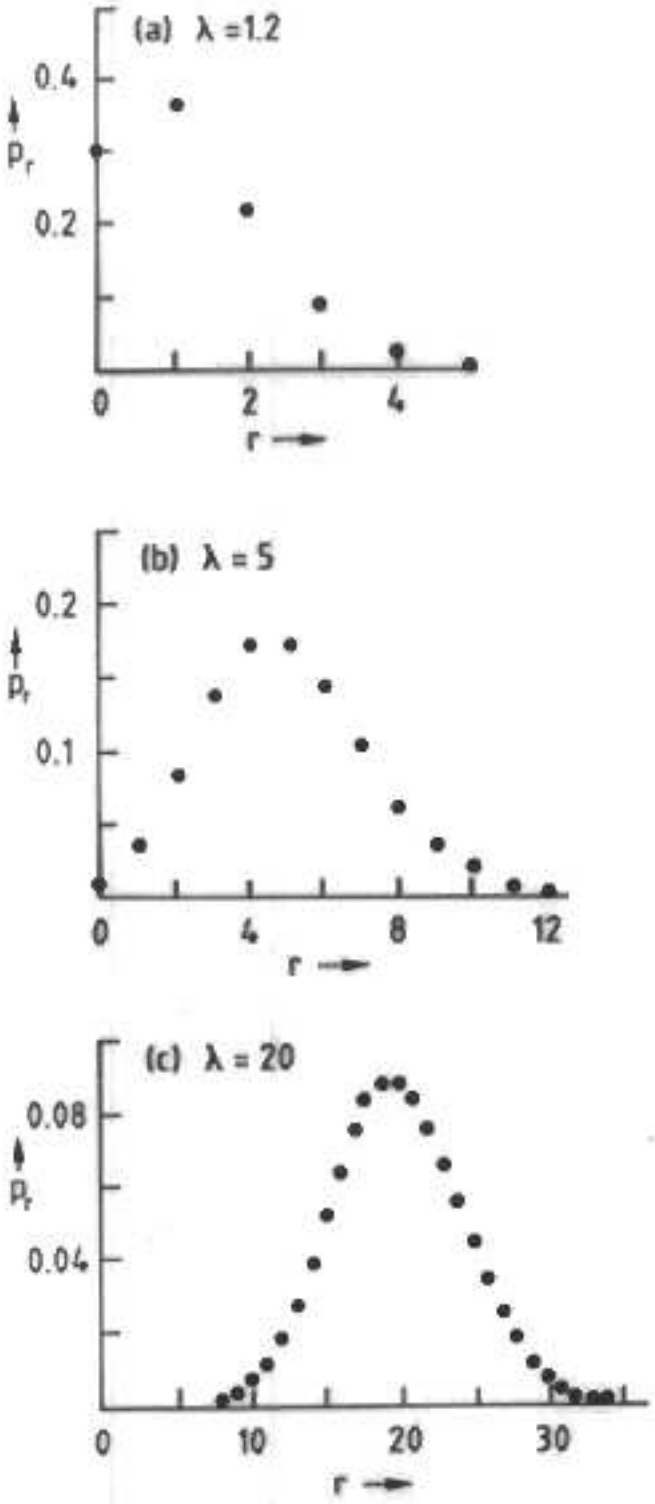
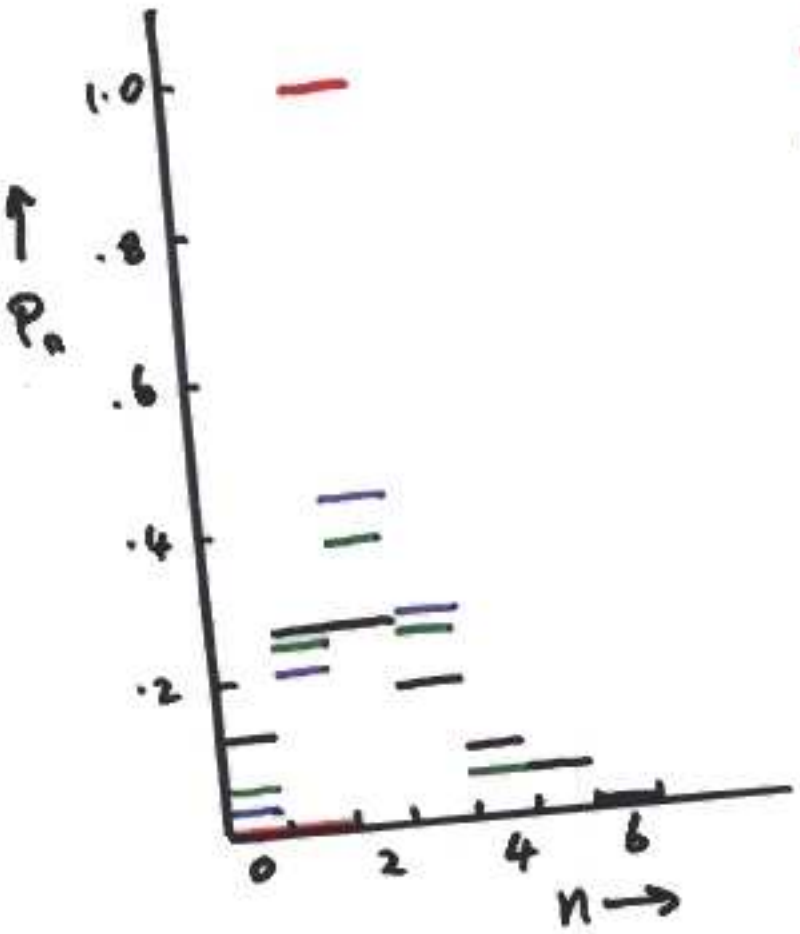


Fig. A4.1 Poisson distributions for different values of the parameter λ. (a) $\lambda = 1.2$; (b) $\lambda = 5.0$; (c) $\lambda = 20.0$. P_r is the probability of observing r events. (Note the different scales on the three figures.) For each value of λ, the mean of the distribution is at λ, and the RMS width is $\sqrt{\lambda}$. As λ increases above about 5, the distributions look more and more like Gaussians.

In a similar way, the Poisson distribution is likely to be applicable to

BINOMIAL ⟹ POISSON



— N = 2 $p = 1$

— N = 3 $p = 2/3$

— N = 4 $p = 1/2$

— Poisson $\mu = 2$

RELATION BETWEEN POISSON AND BINOMIAL

N people at lecture, m males & f females

Assume that these are representative of basic rates :—

ν people ⇑ $\quad \nu p$ males ⇑ $\quad \nu(1-p)$ females

Probability of observing N people

$$P_{Poisson} = \frac{e^{-\nu}\,\nu^{N}}{N!}$$

Probability of given male/female division

$$P_{Binomial} = \frac{N!}{m!\,f!}\,p^{m}(1-p)^{f}$$

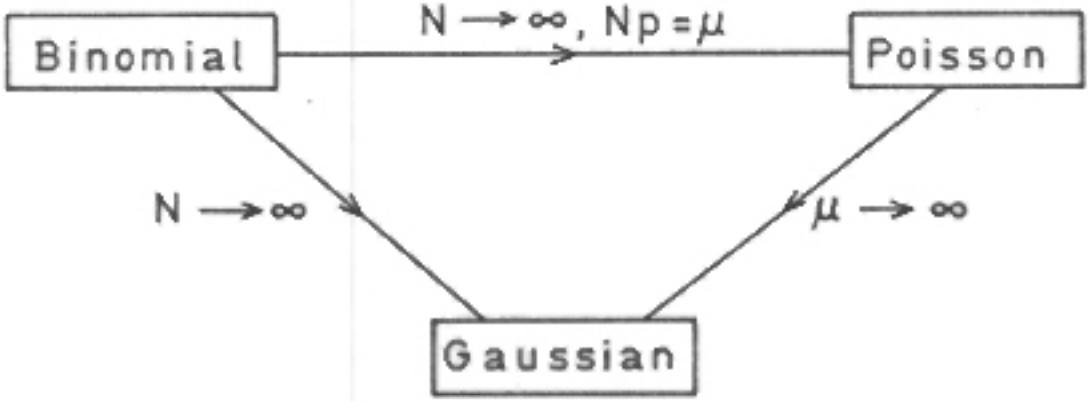
Probability of N people, m males & f females

$$P = P_{Poisson}\,P_{Binomial}$$

$$= \left\{\frac{e^{-\nu p}\,\nu^{m}p^{m}}{m!}\right\} \times \left\{\frac{e^{-\nu(1-p)}\,\nu^{f}(1-p)^{f}}{f!}\right\}$$

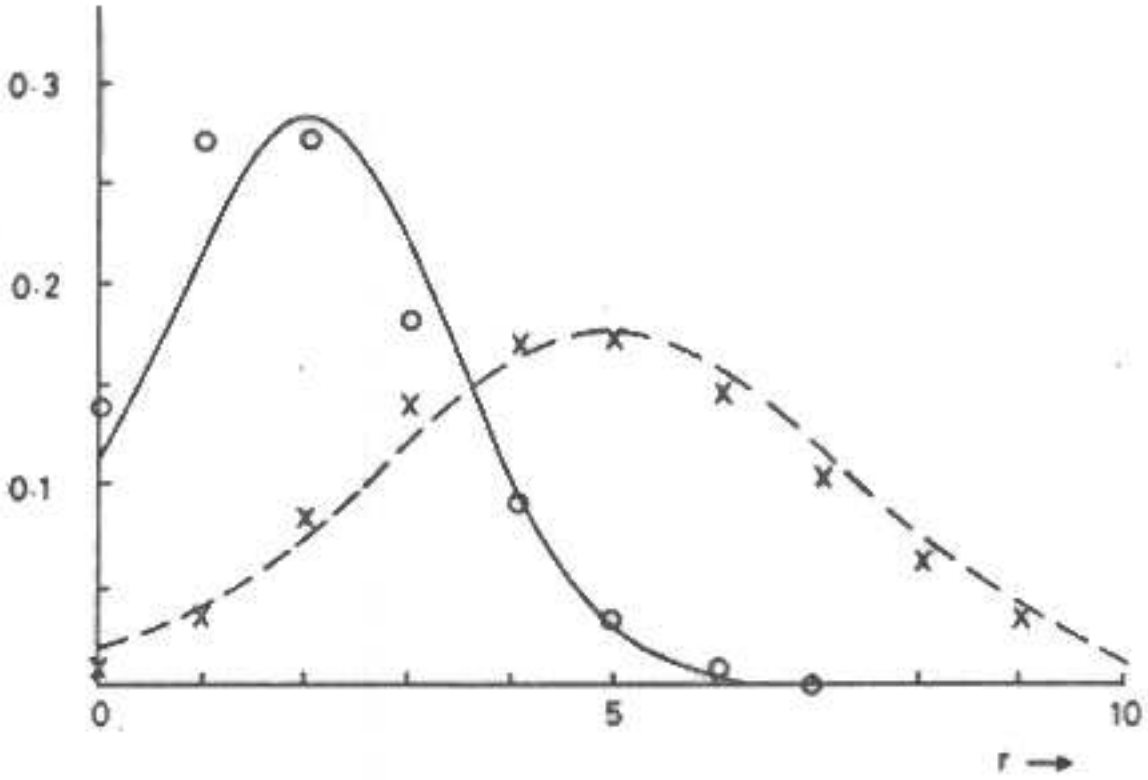
= Poisson distribution for males × Poisson distribution for females

People	Male	Female
Patients	Cured	Remain ill
Decaying nuclei	Forwards	Backwards
Cosmic Rays	Protons	other particles


Binomial
N → ∞, Np = μ
Poisson
N → ∞
μ → ∞
Gaussian

o, x } Poisson

—— , - - - - } Gaussian



Relevant for Hypothesis Testing

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

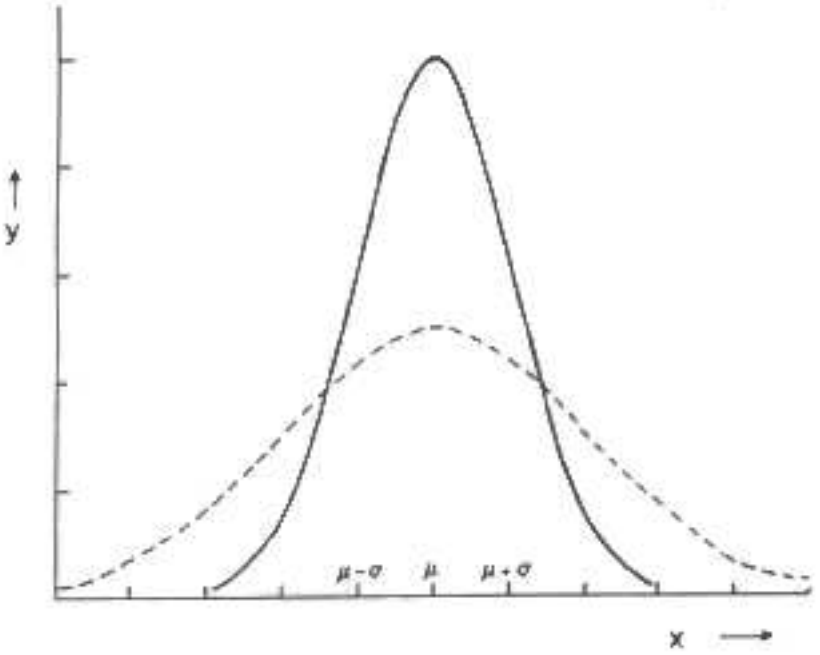


Fig. 1.5. The solid curve is the Gaussian distribution of eqn (1.14). The distribution peaks at the mean μ, and its width is characterised by the parameter σ. The dashed curve is another Gaussian distribution with the same values of μ, but with σ twice as large as the solid curve. Because the normalisation condition (1.15) ensures that the area under the curves is the same, the height of the dashed curve is only half that of the solid curve at their maxima. The scale on the x-axis refers to the solid curve.

Significance of σ

i) RMS of Gaussian $= \sigma$

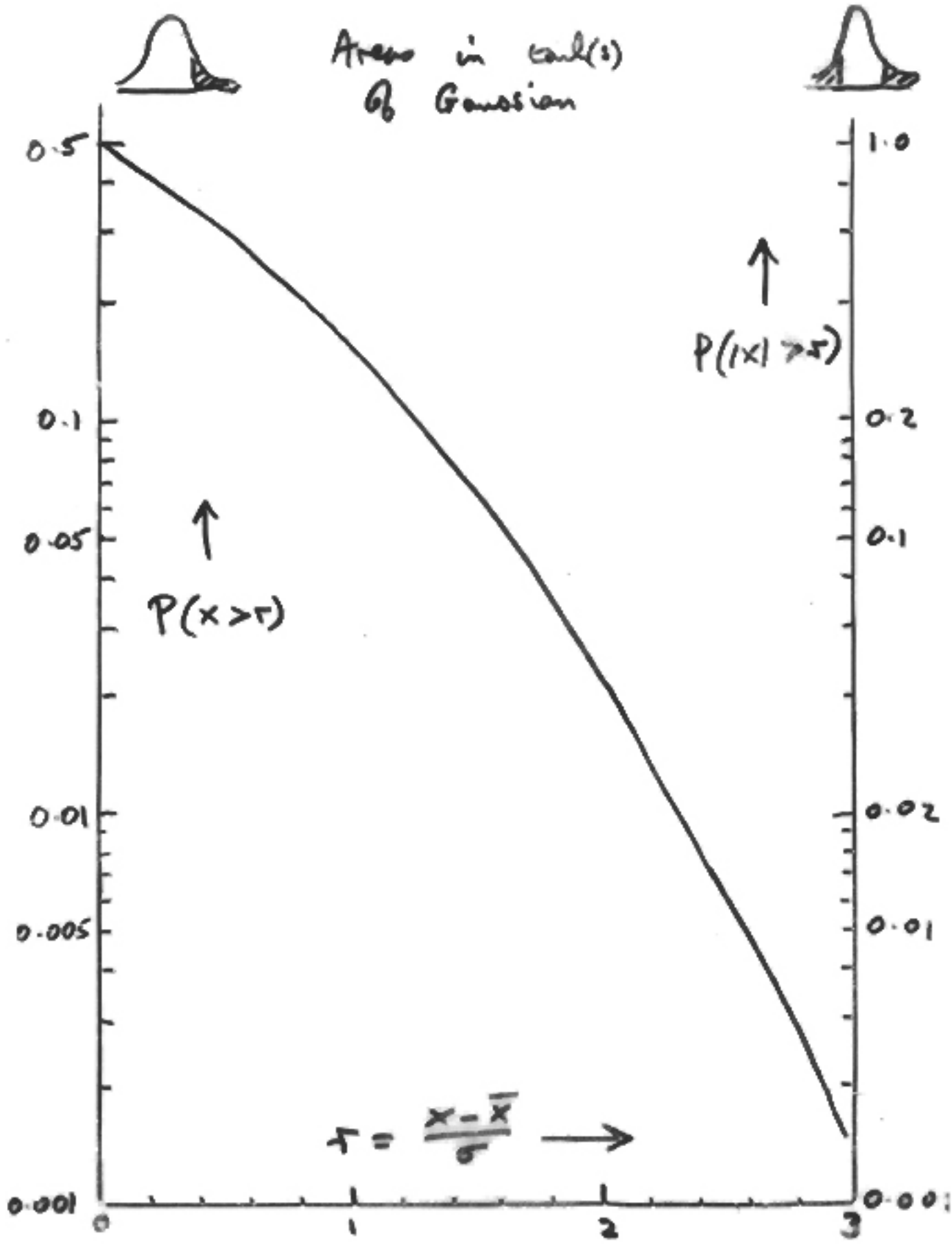
(Hence factor of 2 in defn of Gaussian)

ii) At $x = \mu \pm \sigma$, $y = y_{max}/\sqrt{e}$

(i.e. $\sigma \sim$ half-width at "half" height)

iii) Fractional area within $\mu \pm \sigma$ is 68%

iv) Height at max $= 1/\sqrt{2\pi}\sigma$


Areas in tail(s) of Gaussian
0.5
0.1
0.05
0.01
0.005
0.001
P(x > r)
P(|x| > r)
1.0
0.2
0.1
0.02
0.01
0.00
r = (x - x̄)/σ
0
1
2
3

STUDENT'S t

$x = \frac{\text{meas} - \text{expected}}{\sigma}$

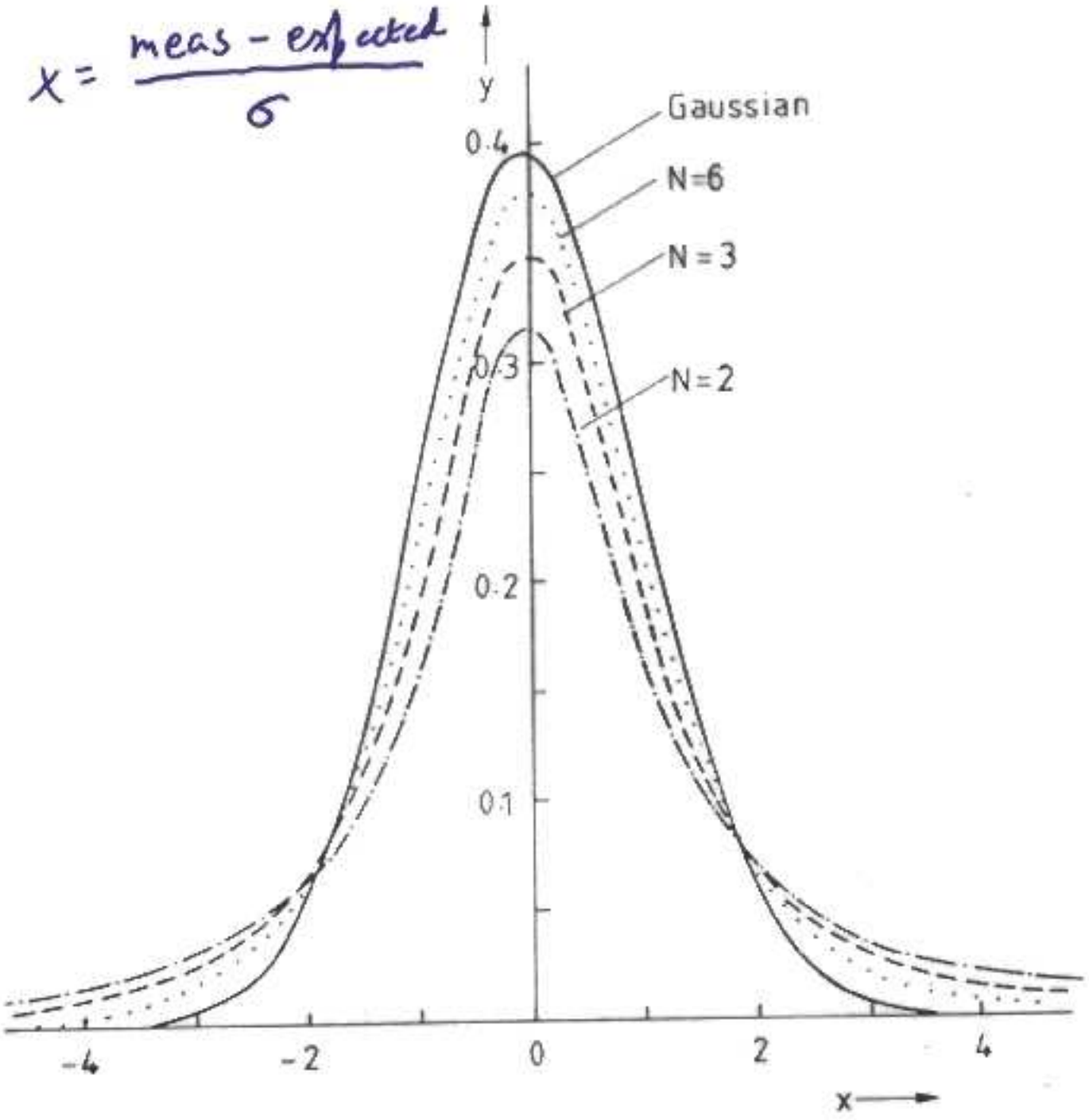


Fig. A5.1 Comparison of Student's t distributions for various values of the number of observations N, with the Gaussian distribution, which is the limit of the Student's distributions as N tends to infinity.

STUDENT'S T

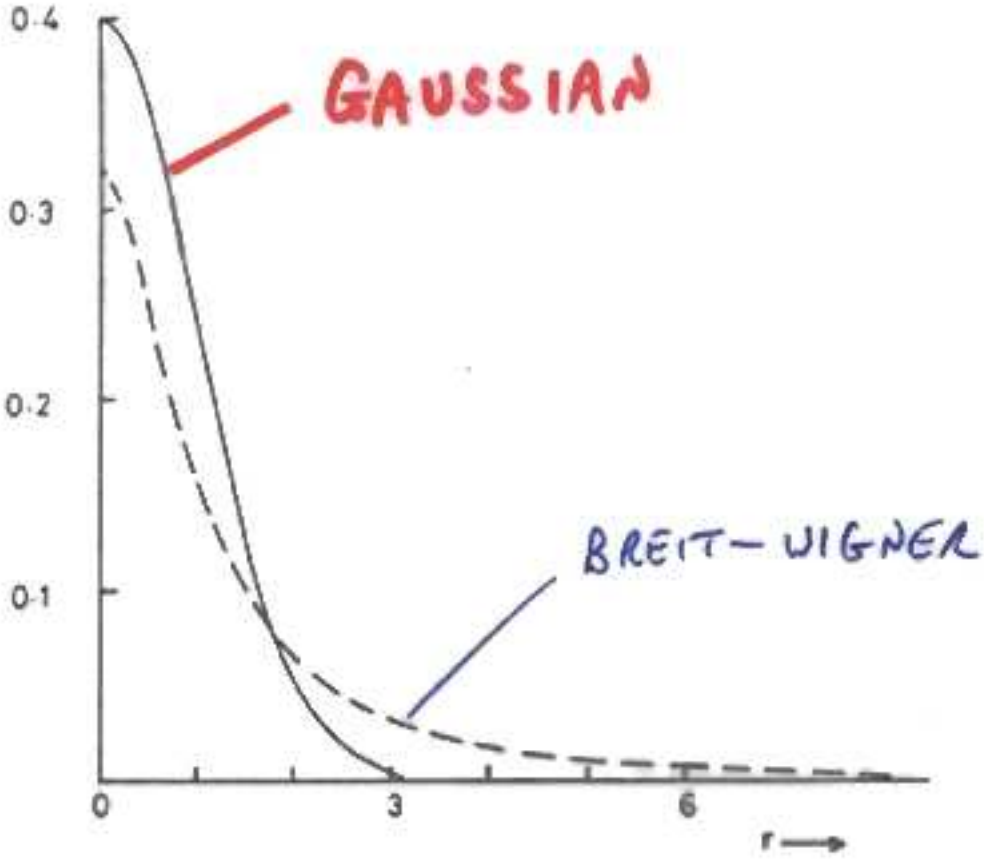
PROB ($t > t_0$)

NDF	0.5%	1%	2.5%	5%	10%	15%
1	64	32	13	6	3.1	1.96
2	10	7	4.3	2.9	1.89	1.37
5	4.0	3.4	2.6	2.0	1.48	1.16
10	3.2	2.8	2.2	1.81	1.37	1.10
30	2.8	2.5	2.0	1.70	1.31	1.06
∞	2.6	2.3	1.96	1.64	1.28	1.04

$$t = \frac{\bar{x} - \mu}{s}$$

Prob ($|t| > t_0$) = 2 * top line

[NDF = ∞] is equivalent to Gaussian.



Gaussian $= N(0,1)$

B-W $= \frac{1}{\pi} \frac{1}{r^2+1}$

Gaussian in 2-variables

$$P(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_x} e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}}$$

$$P(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}}$$

$x + y$ uncorrelated $\Rightarrow$

$$P(x,y) = \frac{1}{2\pi} \frac{1}{\sigma_x \sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)}$$

Down on $P(0,0)$ by $e^{-\frac{1}{2}}$ when

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = 1$$

Rewrite as

$$(x \quad y) \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1$$

Invert $\Rightarrow$ ERROR MATRIX

$$\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

Element E_{ij} is $\langle (x_i - \bar{x}_i)(x_j - \bar{x}_j) \rangle$

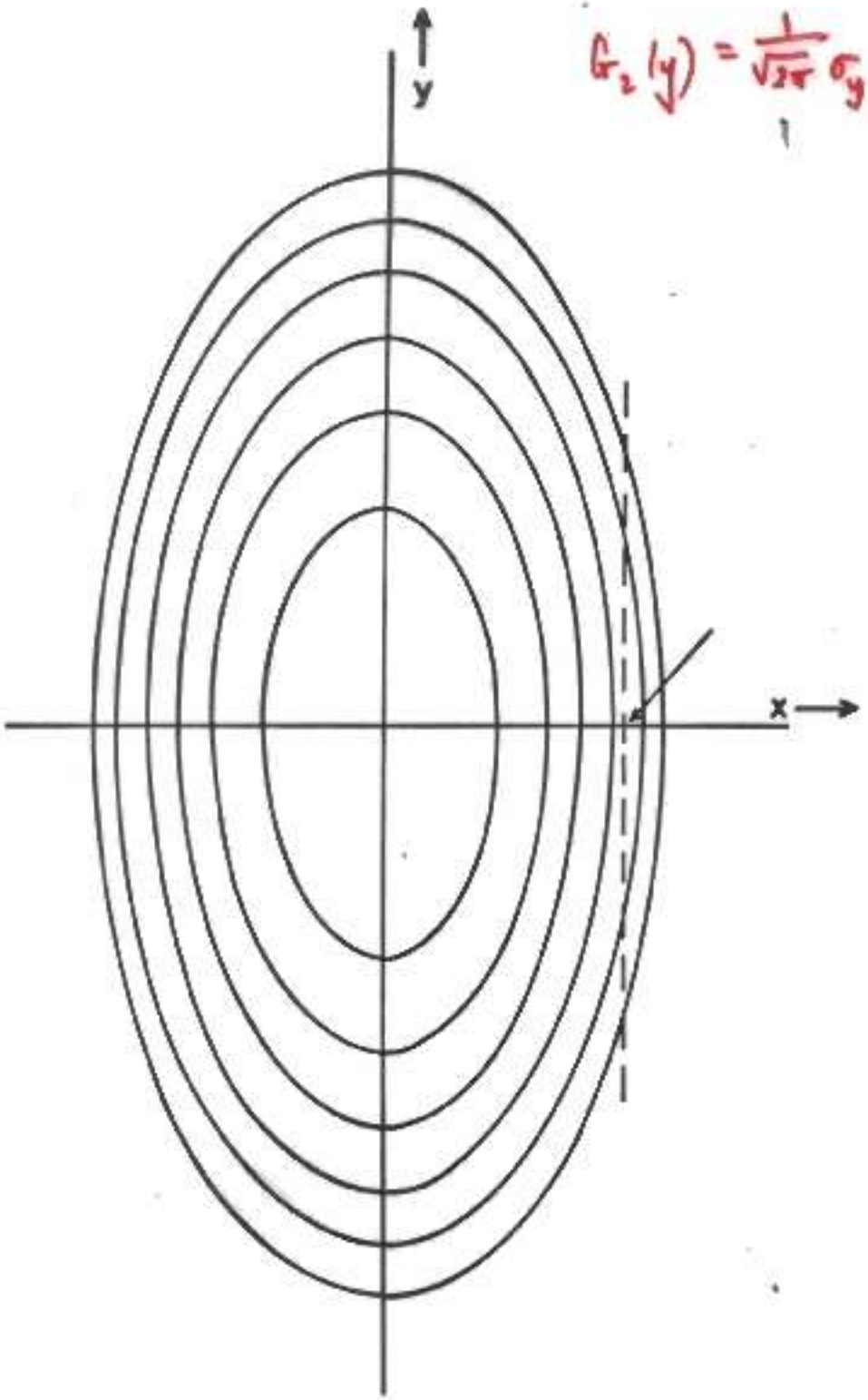
TOWARDS THE ERROR MATRIX

x + y indep Gaussians

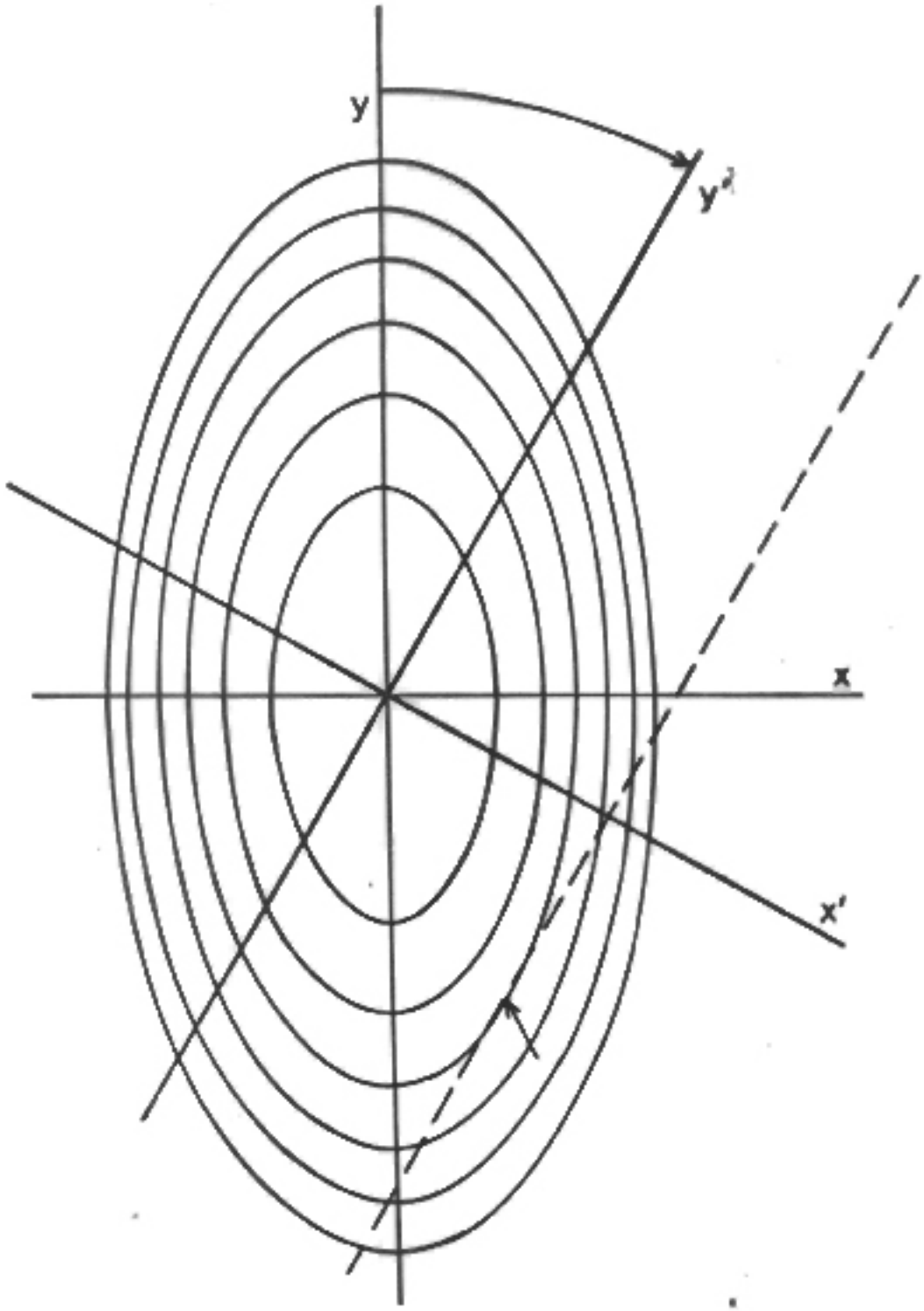
$P(x,y) = G_1(x)\, G_2(y)$

$G_1(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2}\frac{x^2}{\sigma_x^2}\right]$

$G_2(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{1}{2}\frac{y}{\sigma_y^2}\right]$



$$P(x,y) = \frac{1}{2\pi}\frac{1}{\sigma_x^1\sigma_y^1} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right]$$


y
y'
x
x'

Specific example

$$\sigma_x = \frac{\sqrt{2}}{4} = .354 \qquad \sigma_y = \frac{\sqrt{2}}{2} = .707$$

Then factor of $e^{-\frac{1}{2}}$ when

$$8x^2 + 2y^2 = 1$$

Now introduce CORRELATIONS by 30° rotn

$$\tfrac{1}{2}\left[13x'^2 + 6\sqrt{3}\,x'y' + 7y'^2\right] = 1$$

$$\begin{pmatrix} \frac{13}{2} & \frac{3\sqrt{3}}{2} \\ \frac{3\sqrt{3}}{2} & \frac{7}{2} \end{pmatrix} = \text{Inverse Error Matrix}$$

$$\frac{1}{32} \times \begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix} = \text{Error Matrix}$$

$8x^2 + 2y^2 = 1$

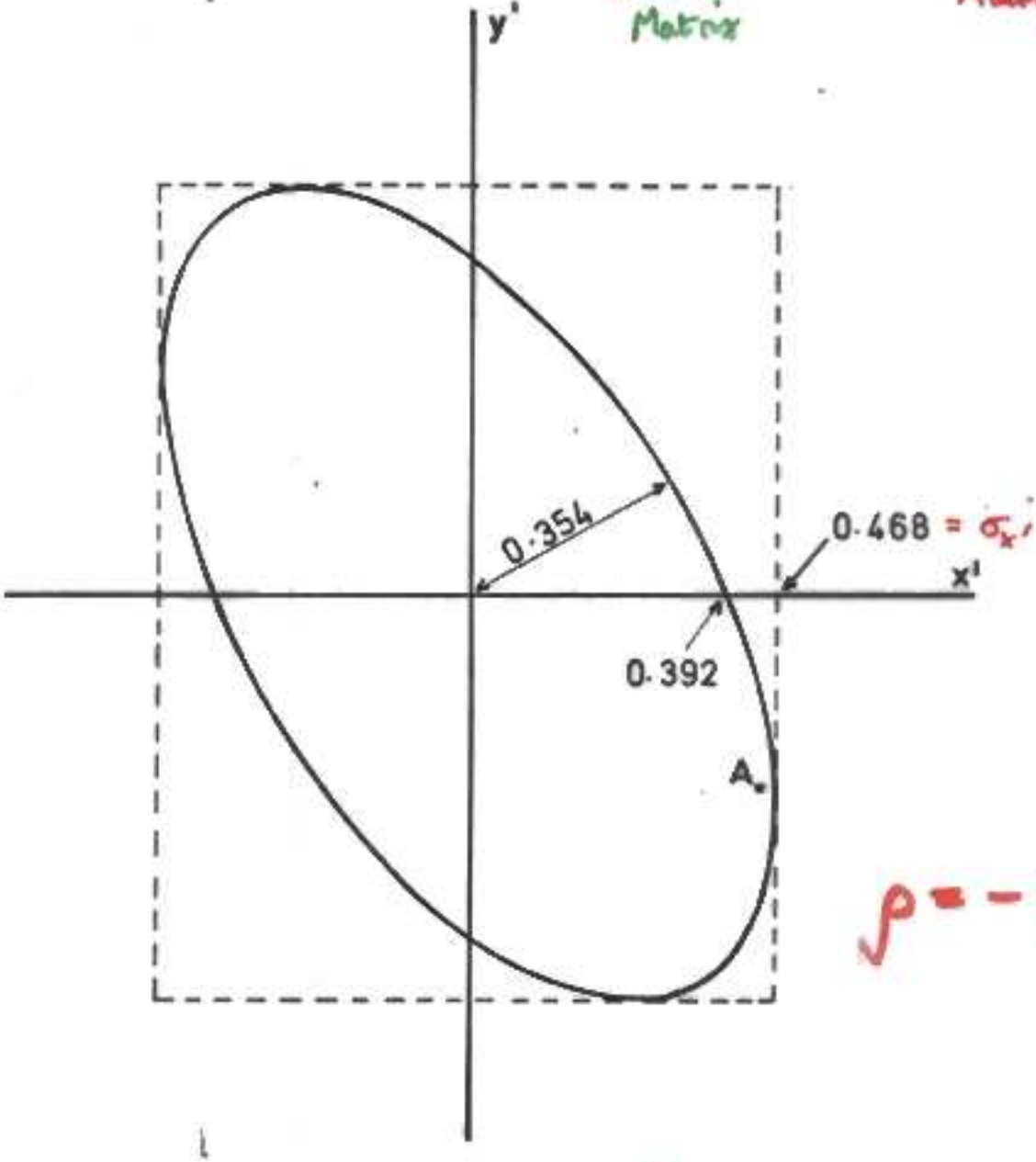
$\frac{1}{2}[13x'^2 + 6\sqrt{3}x'y' + 7y'^2] = 1$

$\begin{pmatrix} \frac{13}{2} & \frac{3\sqrt{3}}{2} \\ \frac{3\sqrt{3}}{2} & \frac{7}{2} \end{pmatrix}$

Inverse Error Matrix

$\frac{1}{32} \times \begin{pmatrix} 7 & -3\sqrt{3} \\ -3\sqrt{3} & 13 \end{pmatrix}$

Error Matrix



$\rho = -0.54$

$(0.468)^2 = \frac{7}{32} = \sigma_{x'}^2$

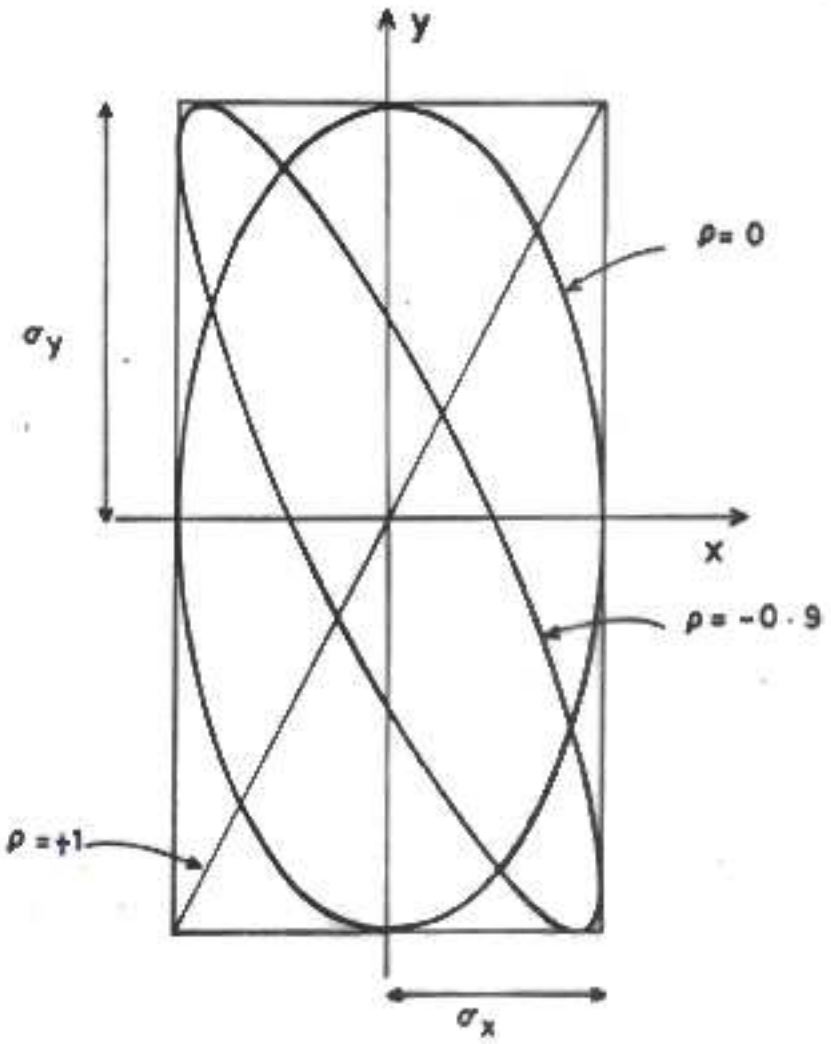
$(0.392)^2 = 1/6.5$

$\frac{1}{8} = (0.354)^2 =$ Eigenvalue of error matrix $= \sigma_x^2$

σ_x, σ_y } constant

ρ varying

Covariance $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$ Error Matrix



USING THE ERROR MATRIX

(i) Function of variables

$$y = y(x_a, x_b)$$

Given x_a, x_b error matrix, what is σ_y?

Differentiate, square, average

$$\overline{\delta y^2} = \left(\frac{\partial y}{\partial x_a}\right)^2 \overline{\delta x_a^2} + \left(\frac{\partial y}{\partial x_b}\right)^2 \overline{\delta x_b^2} + 2 \frac{\partial y}{\partial x_a} \frac{\partial y}{\partial x_b} \overline{\delta x_a \delta x_b}$$

Zero if x_a, x_b uncorrelated

OR

$$\overline{\delta y^2} = \left(\frac{\partial y}{\partial x_a} \quad \frac{\partial y}{\partial x_b}\right) \begin{pmatrix} \overline{\delta x_a^2} & \overline{\delta x_a \delta x_b} \\ \overline{\delta x_a \delta x_b} & \overline{\delta x_b^2} \end{pmatrix} \begin{pmatrix} \frac{\partial y}{\partial x_a} \\ \frac{\partial y}{\partial x_b} \end{pmatrix}$$

$\tilde{D}$ — Error matrix — Derivative vector D

$$\sigma_y^2 = \tilde{D} E D$$

(ii) Change of variables $x_a = x_a(p_i, p_j)$
$x_b = x_b(p_i, p_j)$

e.g. Cartesian ⇒ polars

or Points in x, y ⇒ m, c of straight line fit

Given (p_i, p_j) error matrix ⇒ (x_i, x_j) error matrix

Differentiate, $\delta x_a \delta x_b$, average

$$\delta x_a = \frac{\partial x_a}{\partial p_i}\delta p_i + \frac{\partial x_a}{\partial p_j}\delta p_j \quad (+ \text{sim for } x_b)$$

Then $$\overline{\delta x_a^2} = \left(\frac{\partial x_a}{\partial p_i}\right)^2 \overline{\delta p_i^2} + \left(\frac{\partial x_a}{\partial p_j}\right)^2 \overline{\delta p_j^2} + 2\frac{\partial x_a}{\partial p_i}\frac{\partial x_a}{\partial p_j}\overline{\delta p_i \delta p_j}$$

$$\overline{\delta x_a \delta x_b} = \frac{\partial x_a}{\partial p_i}\frac{\partial x_b}{\partial p_i}\overline{\delta p_i^2} + \frac{\partial x_a}{\partial p_j}\frac{\partial x_b}{\partial p_j}\overline{\delta p_j^2} + \left(\frac{\partial x_a}{\partial p_i}\frac{\partial x_b}{\partial p_j} + \frac{\partial x_a}{\partial p_j}\frac{\partial x_b}{\partial p_i}\right)\overline{\delta p_i \delta p_j}$$

$+ \overline{\delta x_b^2}$ like $\overline{\delta x_a^2}$

N.B. Change of variables does not have to be $N \to N$

e.g. straight line fit involves $N \to 2$

Then i) & ii) are both examples of $N \to M$ $(M \leq N)$

where $M = 1$ in i) $M = N$ in ii)

i.e.

$$\begin{pmatrix} \overline{\delta x_a^2} & \overline{\delta x_a \delta x_b} \\ \overline{\delta x_a \delta x_b} & \overline{\delta x_b^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial x_a}{\partial b_i} & \frac{\partial x_a}{\partial b_j} \\ \frac{\partial x_b}{\partial b_i} & \frac{\partial x_b}{\partial b_j} \end{pmatrix} \begin{pmatrix} \overline{\delta b_i^2} & \overline{\delta b_i \delta b_j} \\ \overline{\delta b_i \delta b_j} & \overline{\delta b_j^2} \end{pmatrix} \begin{pmatrix} \frac{\partial x_a}{\partial b_i} & \frac{\partial x_b}{\partial b_i} \\ \frac{\partial x_a}{\partial b_j} & \frac{\partial x_b}{\partial b_j} \end{pmatrix}$$

↑ New error matrix — ↑ $\tilde{T}$ — ↑ Old error matrix — ↑ Transform matrix T

$$E_x = \tilde{T} E_p T$$

BEWARE!

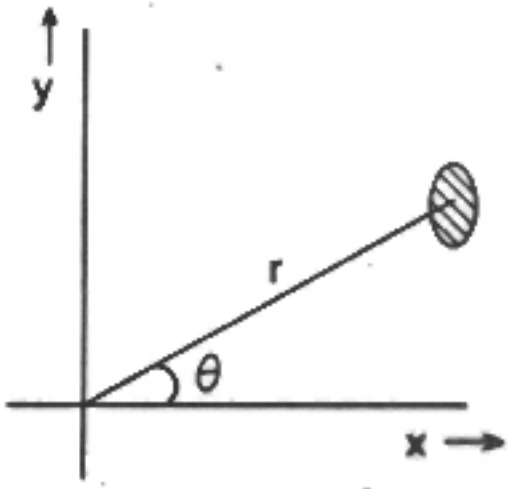
e.g.

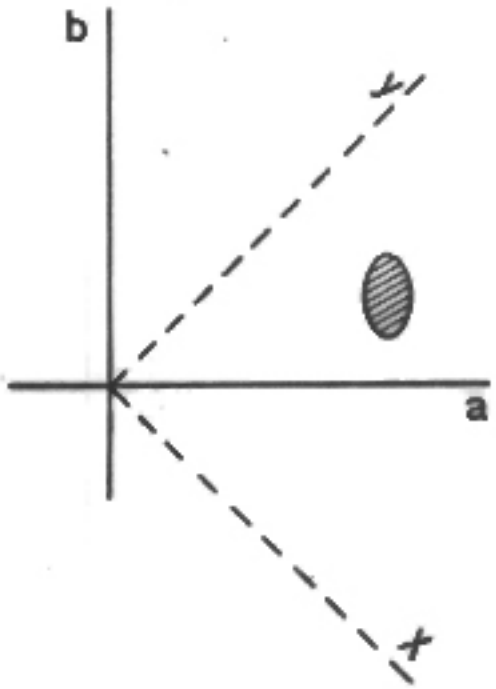
$$\sigma_m^2 = \tilde{D} \tilde{T} E T D$$

Tracks' error matrix (centre of tracks)

Transformation matrix from centre of tracks to vertex

D: vector for mass in terms of track params at vertex


y
r
θ
x


b
a

USING THE ERROR MATRIX

COMBINING RESULTS

If $a_i \pm \sigma_i$ are independent:

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

$$\Rightarrow \hat{a} = \frac{\sum a_i w_i}{\sum w_i} \qquad w_i = 1/\sigma_i^2$$

Now $a_i \pm \sigma_i$ are correlated with error matrix $\underline{\underline{E}}$

$$\underline{\underline{E}} = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(1,2) & \mathrm{cov}(1,3) & \cdots \\ \mathrm{cov}(1,2) & \sigma_2^2 & \mathrm{cov}(2,3) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

$$S = \sum_{i,j} (a_i - \hat{a}) \, \underline{\underline{E}}^{-1}_{ij} \, (a_j - \hat{a})$$

$\underline{\underline{E}}^{-1}$ ← INVERSE ERROR MATRIX

N.B. $\hat{a}$ CAN LIE OUTSIDE a_i

$\sigma_{\hat{a}} \to 0$ AS $\rho \to \pm 1$

$$\underline{\underline{E}}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \cdots \\ 0 & 1/\sigma_2^2 & 0 & \\ \vdots & \vdots & \vdots & \end{pmatrix}$$ FOR UNCORRELATED

MORE COMBINING :

SEVERAL PAIRS OF CORRELATED MEAS.

(x_i, y_i) with $\underline{\underline{E}}_i = \begin{pmatrix} \sigma_x^2 & cov \\ cov & \sigma_y^2 \end{pmatrix}_i$

$$S = \sum_i \left\{ (x_i - \hat{x})^2 E^{-1}_{11,i} + (y_i - \hat{y})^2 E^{-1}_{22,i} + 2(x_i - \hat{x})(y_i - \hat{y}) E^{-1}_{12,i} \right\}$$

ice result: —

Inverse error matrix on result $\hat{x}, \hat{y}$

$$= \sum_i \underline{\underline{E}}_i^{-1}$$

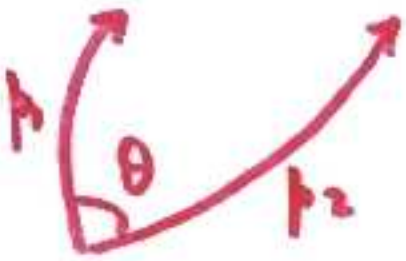
cf $\frac{1}{\sigma^2} = \sum \frac{1}{\sigma_i^2}$ for single uncorrelated meas.

CORRELATIONS + MASS RESOLUTION

$$M^2 = (E_1+E_2)^2 - (\underline{p}_1+\underline{p}_2)^2$$

$$\sim p_1 p_2 \theta \qquad [\, p_i \gg m_i ;\ \theta \ll 1 \,]$$

ie. $M \uparrow$ as $p_i \uparrow$ & $\theta_i \uparrow$



As $p_i \downarrow$, $\theta \uparrow$

$\therefore$ Smaller σ_M

As $p_i \downarrow$, $\theta \downarrow$

$\therefore$ Larger σ_M

ESTIMATING THE ERROR MATRIX

1) ESTIMATE ERRORS

ESTIMATE CORRELATIONS

(Usually easiest if $\rho = 0$ or ± 1)

2) FOR INDEP SOURCES OF ERRORS, ADD ERROR MATRICES

e.g. M_W FROM $WW \to 4$ JETS
$WW \to J J \ell \nu$

$\underline{\underline{E}} = (M_W)_1, (M_W)_2$ ERROR MATRIX

$$\underline{\underline{E}} = \underline{\underline{E}}_{stat} + \underline{\underline{E}}_{B.E.} + \underline{\underline{E}}_{E\ scale} + \underline{\underline{E}}_{FSR} + \underline{\underline{E}}_{colour\ reconn}$$

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \qquad \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix}$$

3) TRANSFORMATIONS

e.g. $(x \pm \sigma_x, y \pm \sigma_y)$ with <u>un</u>correl. errors

$\Rightarrow$ r, θ <u>with</u> correlations

Indep data points

$\Rightarrow$ correlated a and b

Track fit

4) REPEATED OBSERVATIONS

$(x_i, y_i) \Rightarrow \sigma_x^2 \quad \sigma_y^2$ and

$cov(x, y)$ from $\overline{(x-\bar{x})(y-\bar{y})}$

FNAL STATISTICS LECTURE #2

LOUIS LYONS

PARAMETER DETERMINATION

METHOD OF MOMENTS

MAXIMUM LIKELIHOOD

(LEAST SQUARES NEXT TIME)

MOMENTS

e.g. $\frac{dn}{d\cos\theta} = N\left(1 + \frac{b}{a}\cos^2\theta\right)$

$$\overline{\cos^2\theta} = \frac{\frac{1}{3} + \frac{1}{5}\frac{b}{a}}{1 + \frac{1}{3}\frac{b}{a}}$$

$$\Rightarrow \frac{b}{a} = \frac{5\left(3\overline{\cos^2\theta} - 1\right)}{\left(3 - 5\overline{\cos^2\theta}\right)}$$

Check $\overline{\cos^2\theta} = \frac{1}{3} \Rightarrow \frac{b}{a} = 0$

ie. $\frac{b}{a}$ from averaging $\cos^2\theta$ over events

Error from error on $\overline{\cos^2\theta}$ by

i) spread of $\cos^2\theta_i$

OR ii) from expected distribution ← Better

[Error on $\frac{b}{a}$ asymmetric]

MOMENTS, contd

1) Very easy. No maximisation

2) No binning needed

3) Extendable to several parameters/variables

4) Constraints among params not readily incorporated

5) Params can be unphysical

e.g. $\overline{\cos^2\theta} = 0 \Rightarrow \frac{b}{a} = -5/3$

$\overline{\cos^2\theta} = 1 \Rightarrow \frac{b}{a} = -5$

$\overline{\cos^2\theta} = 3/5 + \varepsilon \Rightarrow \frac{b}{a} \sim -\infty$

6) Several possible moments

e.g. $\overline{\cos^4\theta}$ $\overline{|\cos\theta|}$

7) No check on "goodness of fit"

Useful ⇒ starting values for more sophisticated methods

MAXIMUM LIKELIHOOD

WHAT IT IS

HOW IT WORKS : RESONANCE

ERROR ESTIMATES :
STATUS OF $\Delta \ln L = -0.5$ RULE

DETAILED EXAMPLE : LIFETIME

LIKELIHOOD → PDF : TRANSFORMATION PROPS. OF $\mathcal{L}$

SEVERAL PARAMETERS

LIKELIHOOD → GOODNESS OF FIT

BINNED → UNBINNED $\mathcal{L}$

EXTENDED MAX LIKE

COMMENTS ON $\mathcal{L}$ METHOD

BIAS FROM INCORRECT $\mathcal{L}$

BAYESIAN SMEARING OF $\mathcal{L}$ OR $\ln \mathcal{L}$

MAXIMUM LIKELIHOOD

$$y = N\left(1 + \frac{b}{a}\cos^2\theta\right)$$

$$y_i = N\left(1 + \frac{b}{a}\cos^2\theta_i\right)$$

~ Probability density of observing θ_i, given b/a

$$\mathcal{L}\left(\frac{b}{a}\right) = \prod y_i$$

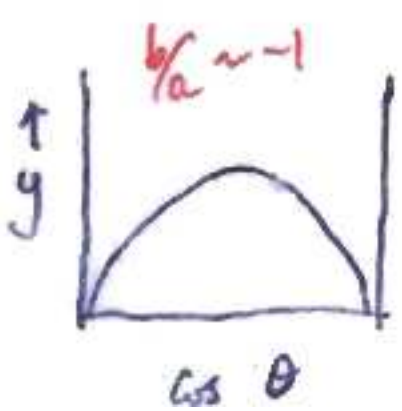
~ Probability of observing given set of θ_i for that b/a

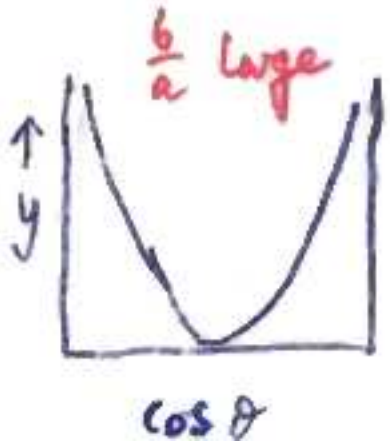
Best estimate of $\frac{b}{a}$ is that which maximises $\mathcal{L}$

Precision of $\frac{b}{a}$ from width of $\mathcal{L}$ distribution

CRUCIAL TO NORMALISE y

SHAPE DETERMINES PARAMS





1) NORMALISATION OF $\mathcal{L}$

$\int P(x|\mu)\,dx$ MUST BE INDEPENDENT OF μ

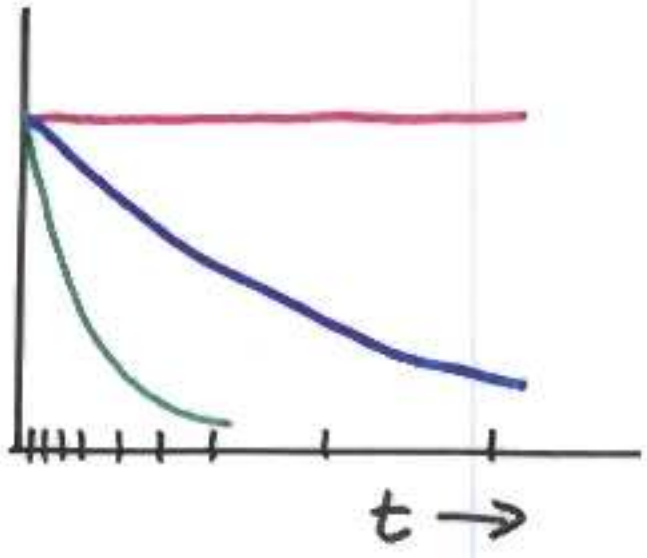
DATA PARAM

e.g. Lifetime fit to $t_1, t_2, \ldots t_n$

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

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

↑ Missing $1/\tau$



— $\tau = \infty$

— τ too big

— Reasonable τ

1) NORMALISATION OF $\mathcal{L}$

$\int P(x|\mu)\,dx$ MUST BE INDEPENDENT OF μ

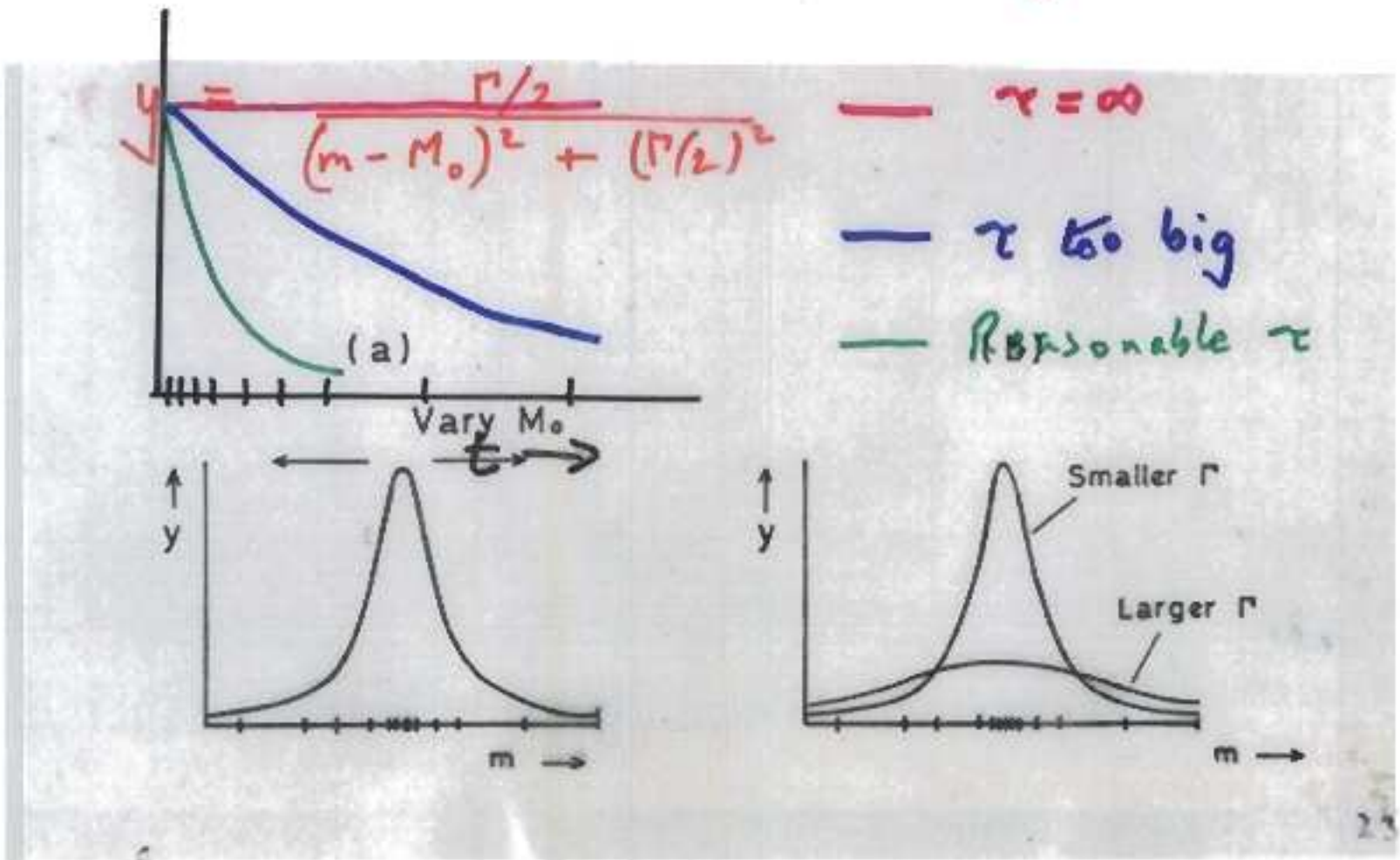
DATA PARAM

e.g. Lifetime fit to $t_1, t_2, \ldots t_n$

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

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

↑ Missing $1/\tau$



Conventional to consider

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

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

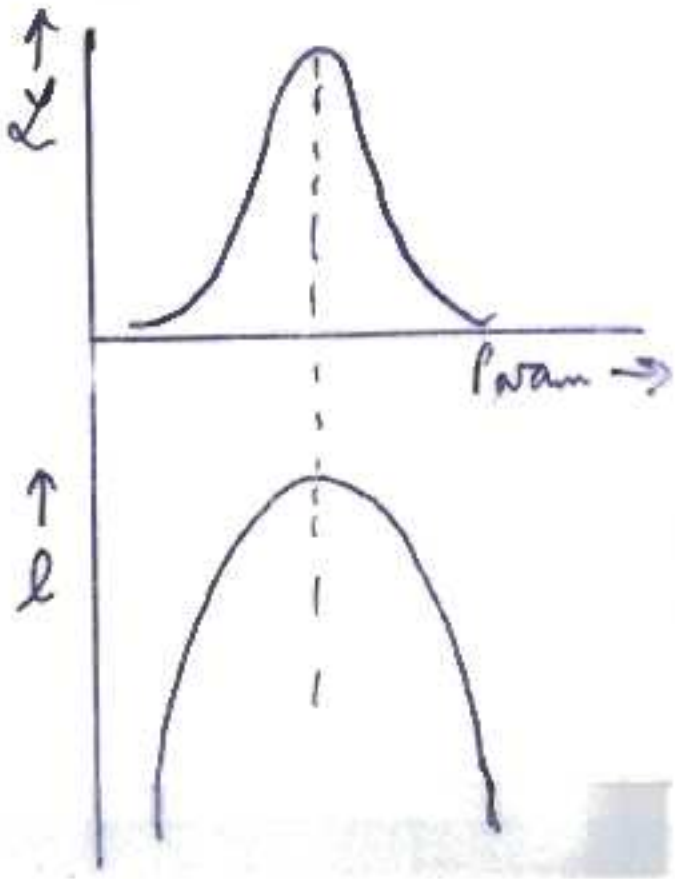
"Proof"

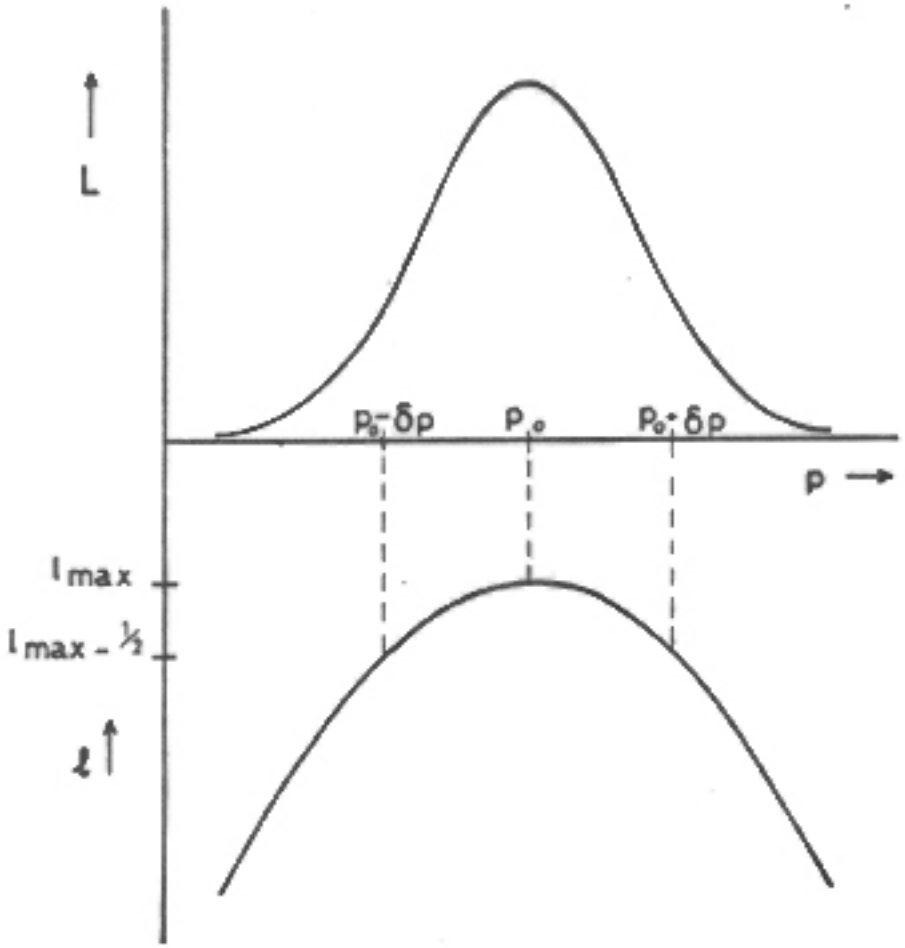
Taylor expand l about its maximum

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

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

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




L
$P_0 - \delta p$
P_0
$P_0 + \delta p$
p
I_{max}
$I_{max} - \frac{1}{2}$
I

MAXIMUM LIKELIHOOD ERROR

Range of likely values of p = param from width of $\mathcal{L}$ or l distributions

When $\mathcal{L}$ is Gaussian, following are equiv

1) RMS of $\mathcal{L}$ distribution

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

3) Change in p so that $l = l_{max} - \frac{1}{2}$

If $\mathcal{L}$ is non-Gaussian, 3) still gives range of p that corresponds to 68% prob (though not usually the shortest)

Error usually asymmetric.

Asymmetric errors are messy, so try to choose parameters intelligently

e.g λ or τ

b or $1/b$

COVERAGE:

HOW OFTEN DOES QUOTED RANGE FOR PARAM INCLUDE PARAM'S TRUE VALUE

N.B. COVERAGE IS PROPERTY OF METHOD, NOT OF A PARTICULAR LIMIT

COVERAGE CAN VARY WITH μ

—"—

STUDY COVERAGE OF DIFFERENT METHODS OF POISSON PARAMETER μ FOR OBSERVATION OF NUMBER OF EVENTS n

HOPE FOR:

COVERAGE

If true for all μ : "correct coverage"

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

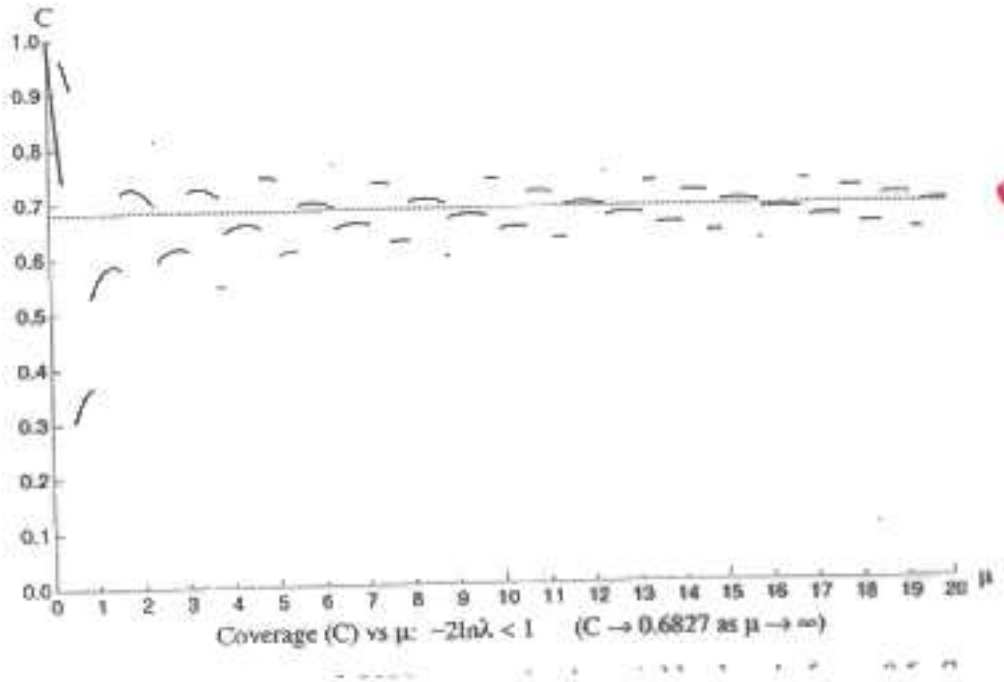
P> α for some μ: "overcoverage"

Conservative

Loss of rejection power

L approach

NOT frequentist



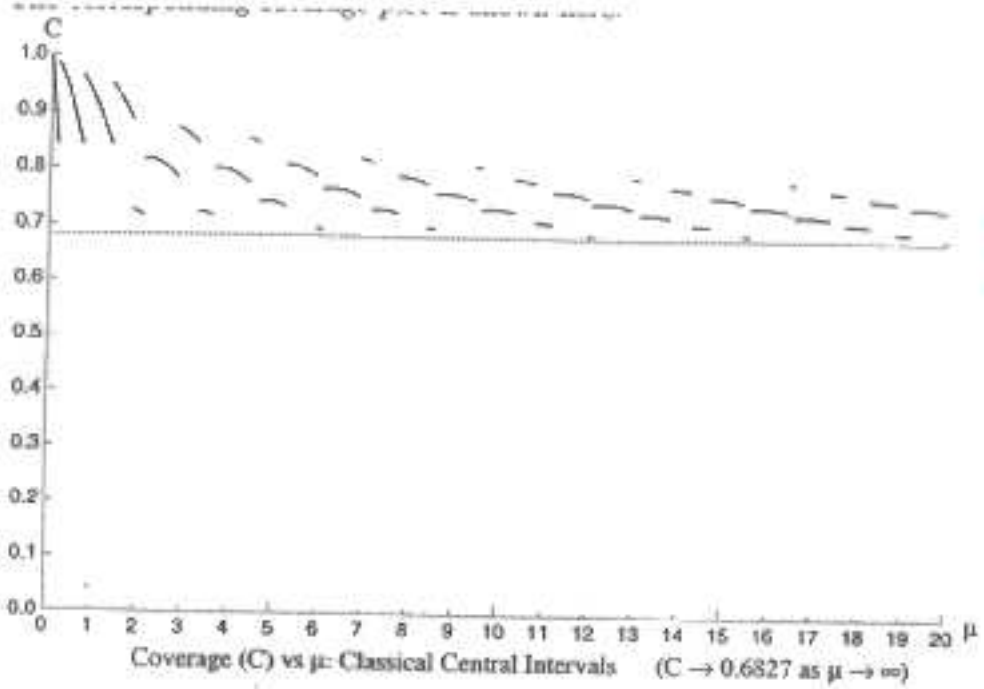
$$P(n,\mu) = e^{-\mu}\mu^n / n!$$

$$-2 \ln \lambda < 1$$

$$\left[\lambda = P(n,\mu) \,/\, P(n,\mu_{best})\right]$$

COVERAGE OF ERROR BARS FOR POISSON DATA

JOEL HEINRICH
CDF 6438



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

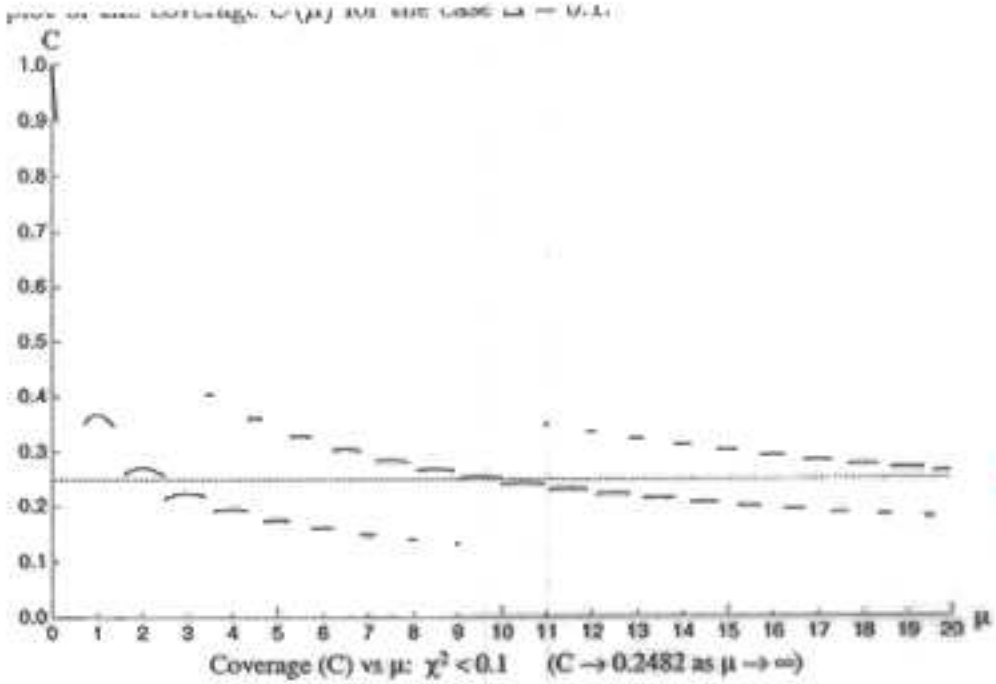
Classical central intervals

at 68.3% coverage

Never undercovers

(conservative at both ends)

COVERAGE OF ERROR BARS FOR POISSON DATA — JOEL HEINRICH CDF 6438



3

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

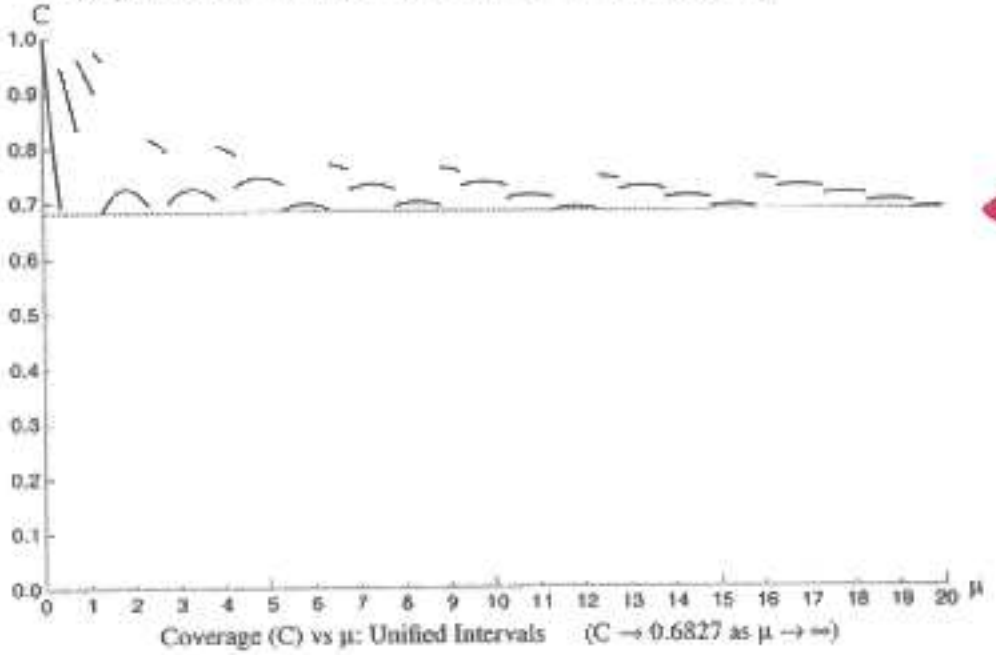
$$\chi^2 = \left(\frac{n-\mu}{\sqrt{\mu}}\right)^2$$

$$\Delta\chi^2 = 0.1 \Longrightarrow 24.8\% \text{ coverage}$$

χ^2 Approach

NOT FREQUENTIST

Finally, we show the coverage of the 1σ unified intervals:

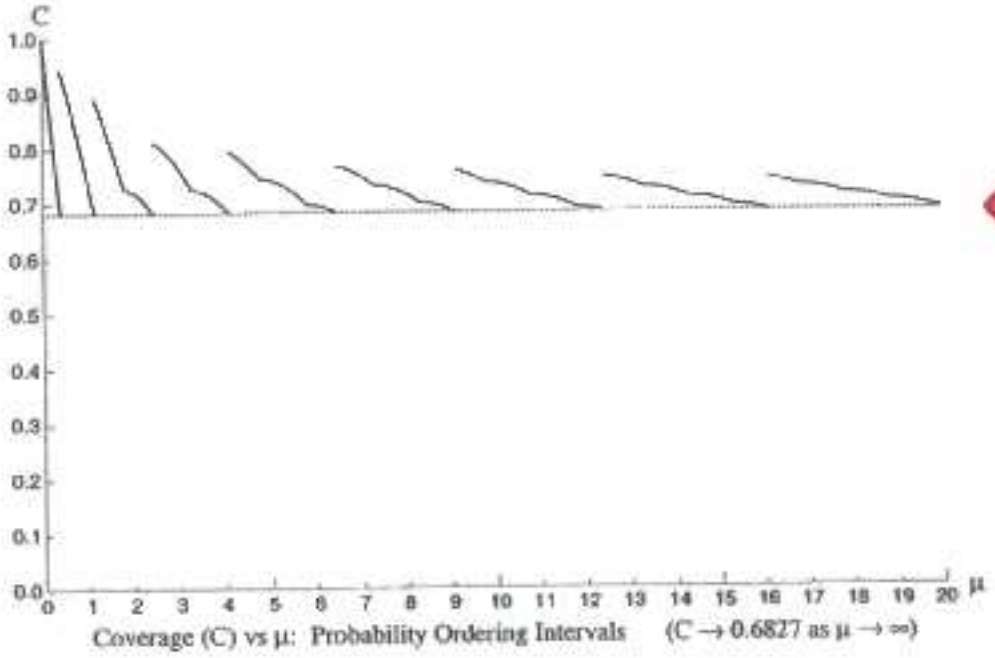


Coverage (C) vs μ: Unified Intervals ($C \to 0.6827$ as $\mu \to \infty$)

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

Unified intervals at 68.3% coverage

↳ Feldman Cousins



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

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

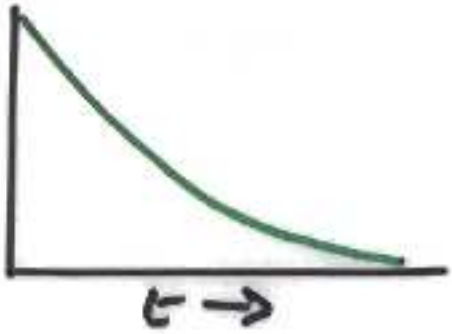
Probability ordering intervals

at 68.3% coverage

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

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

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

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

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

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

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

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

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

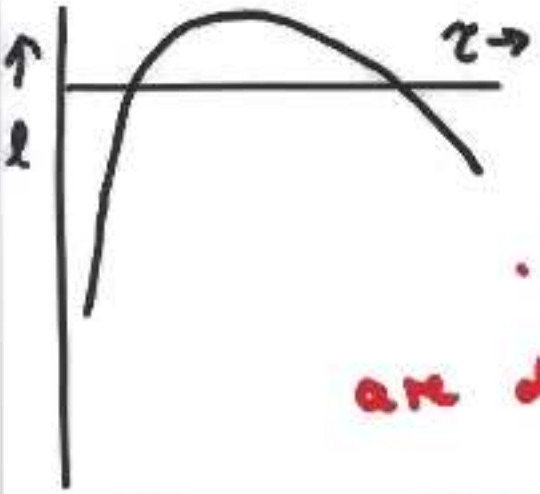
BEWARE FOR AVERAGING RESULTS

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

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

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

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



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

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

— " —

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

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

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

$\mathcal{L}$ AND pdf

EXAMPLE 1 POISSON

pdf = Probability distribution function for observing n, given μ, is

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

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

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

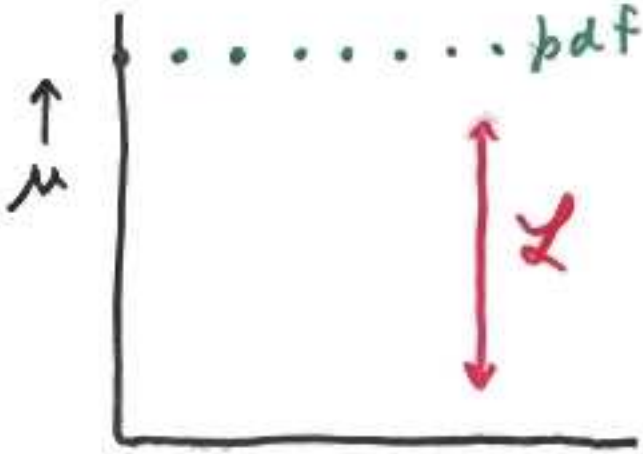
i.e. Use same function of μ & n, but

for pdf, μ is fixed

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

N.B. $P(n;\mu)$ exists only at integer $n \geq 0$

$\mathcal{L}(\mu;n)$ exists as continuous fn of $\mu \geq$



Example 2 Lifetime distribution

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

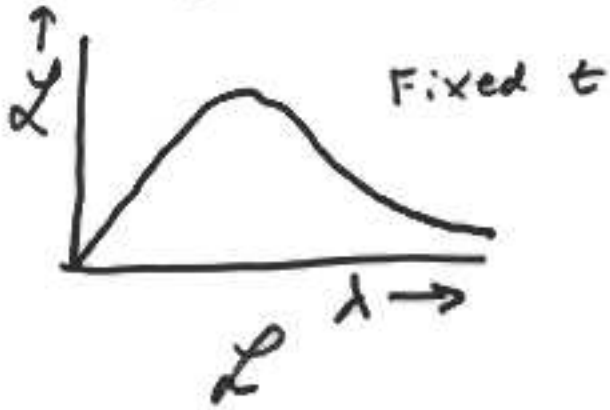
$\therefore \mathcal{L}(\lambda;t) = \lambda e^{-\lambda t}$ [Single observed t]

Now both t & λ are continuous

pdf maximises at $t=0$

$\mathcal{L}$ ..—.—.— $\lambda = t$

pdf



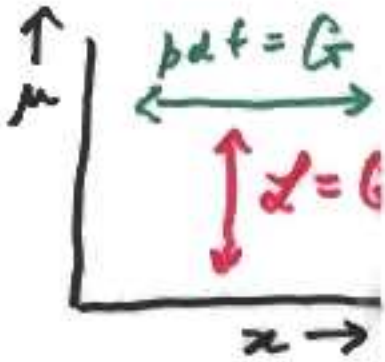
$\mathcal{L}$

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

EXAMPLE 3 GAUSSIAN

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

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



N.B. In this case, same functional form

$\therefore$ If only consider Gaussians, can get confused between pdf & $\mathcal{L}$

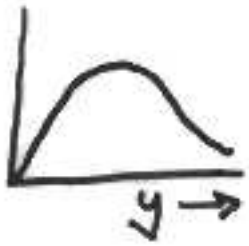
$\therefore$ Examples 1 & 2 are useful

TRANSFORMATION PROPERTIES OF $\mathcal{L}$ AND PROBABILITY DENSITIES

LIFETIME EXAMPLE $\frac{dn}{dt} = \lambda e^{-\lambda t}$

CHANGE <u>OBSERVABLE</u> FROM t to $y = +\sqrt{t}$

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



So (a) pdf changes, BUT

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

CONTRAST $\mathcal{L}$, WHICH IS <u>NOT</u> pdf for λ, GIVEN observed t

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

(a') $\mathcal{L}$ DOES NOT CHANGE

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

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

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

BUT (b') $\int_{\lambda_0}^{\infty} \mathcal{L}(\lambda; t)\,d\lambda \neq \int_{\tau_0 = 1/\lambda_0}^{\infty} \mathcal{L}(\tau; t)\,dt$

∴ It is <u>not</u> meaningful to integrate $\mathcal{L}$

	$pdf(t;\lambda)$	$\mathcal{L}(\lambda;t)$
VALUE OF FUNCTION	CHANGES WHEN OBSERVABLE IS TRANSFORMED	INVARIANT WRT TRANSFORMATION OF PARAMETER
INTEGRAL OF FN	INVARIANT W.R.T. TRANSFORMATION OF OBSERVABLE	CHANGES WHEN PARAM IS TRANSFORMED
CONCLUSION	MAX PROB DENSITY NOT VERY SENSIBLE	INTEGRATING $\mathcal{L}$ NOT VERY SENSIBLE

CONCLUSION:

$$\int_{p_L}^{p_U} \mathcal{L}\, dp = \alpha$$ " 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 $\mathcal{L} \Rightarrow$ BAYES POSTERIOR

$$P_{posterior}(\lambda) \propto \mathcal{L}(\lambda) \times \pi(\lambda)$$

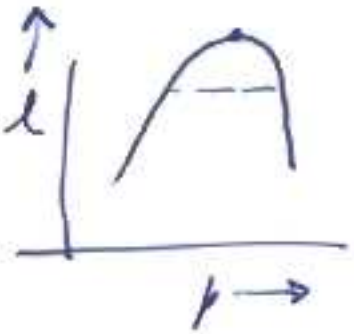
$\pi(\lambda)$ would be const

SEVERAL PARAMETERS

1 param p

p from $\frac{\partial l}{\partial p} = 0$

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



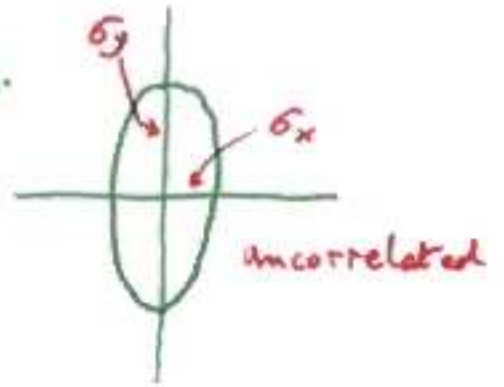
Many dimensions : $l(p_1, p_2, p_3 \ldots)$

$p_1\ p_2\ p_3 \ldots$ from $\frac{\partial l}{\partial p_i} = 0$

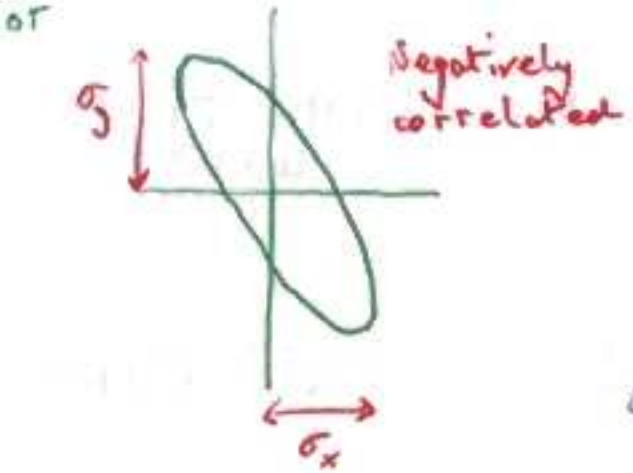
For errors, define $H_{ij} = -\frac{\partial^2 l}{\partial p_i \partial p_j}$ = Inverse Error Matrix

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

e.g.



or



N.B. ERROR NOT GIVEN BY

$l = l_{max} - \frac{1}{2}$ WHEN VARYING x

FROM BEST VALUE WHILE

KEEPING y.... CONSTANT

ERROR IS GIVEN BY

$l = l_{max} - \frac{1}{2}$ WHEN VARYING x

FROM BEST VALUE WHILE .....

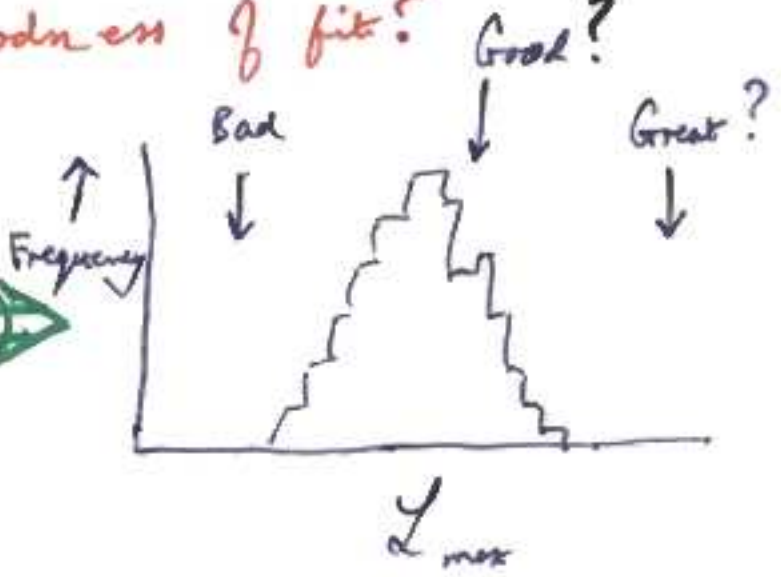
$\mathcal{L}_{max}$ and Goodness of Fit ?

Find parameters by maximising $\mathcal{L}$

$\therefore$ Larger $\mathcal{L}_{max}$ better than smaller $\mathcal{L}_{max}$

$\therefore$ Use $\mathcal{L}_{max}$ $\Rightarrow$ Goodness of fit?

M.C. distribution of unbinned $\mathcal{L}_{max}$



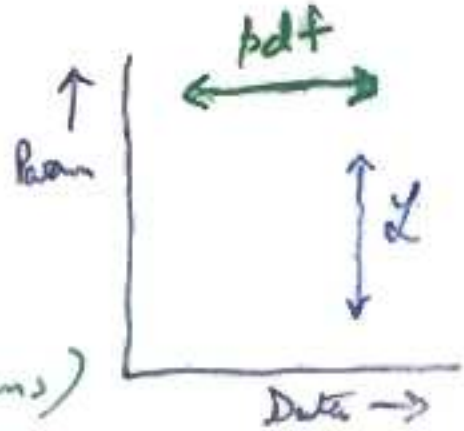
Not necessarily:

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

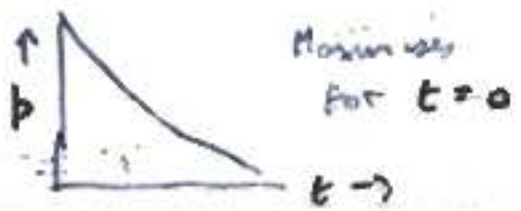
↑ Fixed ↑ VARY

Contrast pdf = p (data , params)

↑ VARY ↑ FIXED



e.g. $p(t,\tau) = \lambda e^{-\lambda t}$



Examples

① Fit exponential to times $t_1, t_2, t_3 \ldots$

[Joel Heinrich, CDF]

$$\mathcal{L} = \prod_i \lambda e^{-\lambda t_i}$$

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

i.e. Depends only on $\bar{t}$, a is

INDEPENDENT OF DISTRIBUTION OF t

(except for ....)

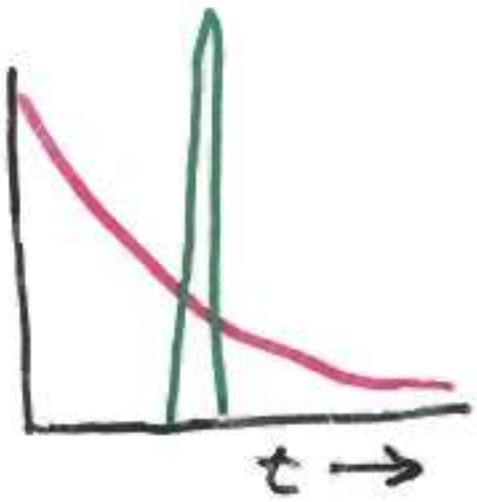
[$\bar{t}$ is SUFFICIENT STATISTIC]

∴ variation of $\mathcal{L}_{max}$ in M.C. due to variations in sample $\bar{t}$

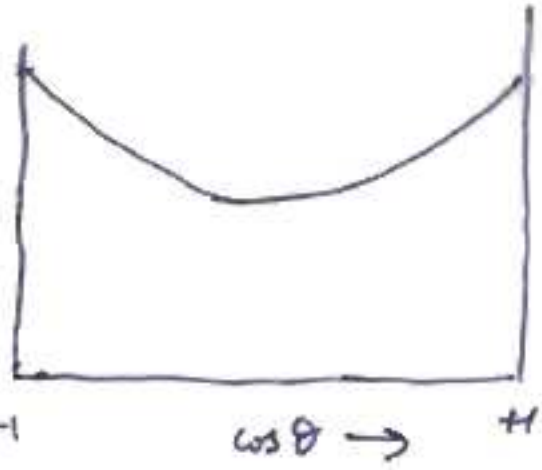
NOT TO BETTER OR WORSE FIT

Same $\bar{t}$

⇒ same $\mathcal{L}_{max}$ →



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



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

DEPENDS ONLY ON $\cos^2\theta_i$:

INSENSITIVE TO SIGN OF $\cos\theta_i$

∴ Data can be in very bad agreement with expected distribution (e.g. all data with $\cos\theta > 0$), & $\mathcal{L}_{max}$ does not know it.

Example of general principle.

3) Fit data to Gaussian, with variable μ
fixed σ

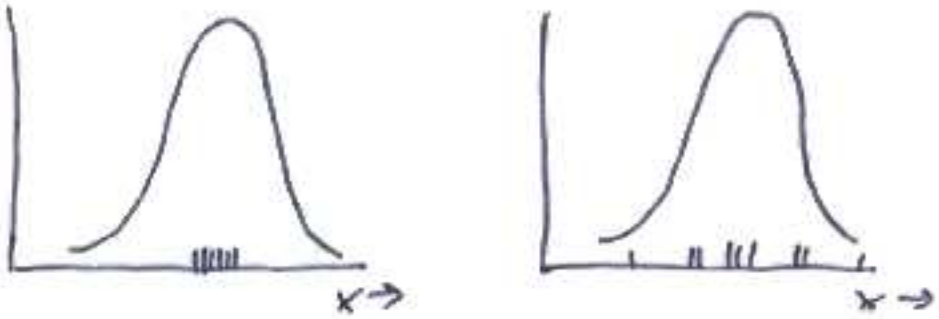
$$\text{pdf} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\tfrac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$\ln \mathcal{L}_{max} = N\underbrace{\left(-\tfrac{1}{2}\ln 2\pi - \ln\sigma\right)}_{\text{const}} - \tfrac{1}{2}\underbrace{\sum (x_i-\bar{x})^2/\sigma^2}_{\frac{\text{Variance}(x)}{\sigma^2}}$$

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

which isn't even relevant for fitting μ $(\hat{\mu} = \bar{x})$

Smaller than expected variance $\Rightarrow$ larger $\mathcal{L}_{max}$



WORSE FIT
LARGER $\mathcal{L}_{max}$

BETTER FIT
LOWER $\mathcal{L}_{max}$

$\mathcal{L}$ has sensible properties wrt parameters
NOT wrt data

$\mathcal{L}_{max}$ & GOODNESS OF FIT ?

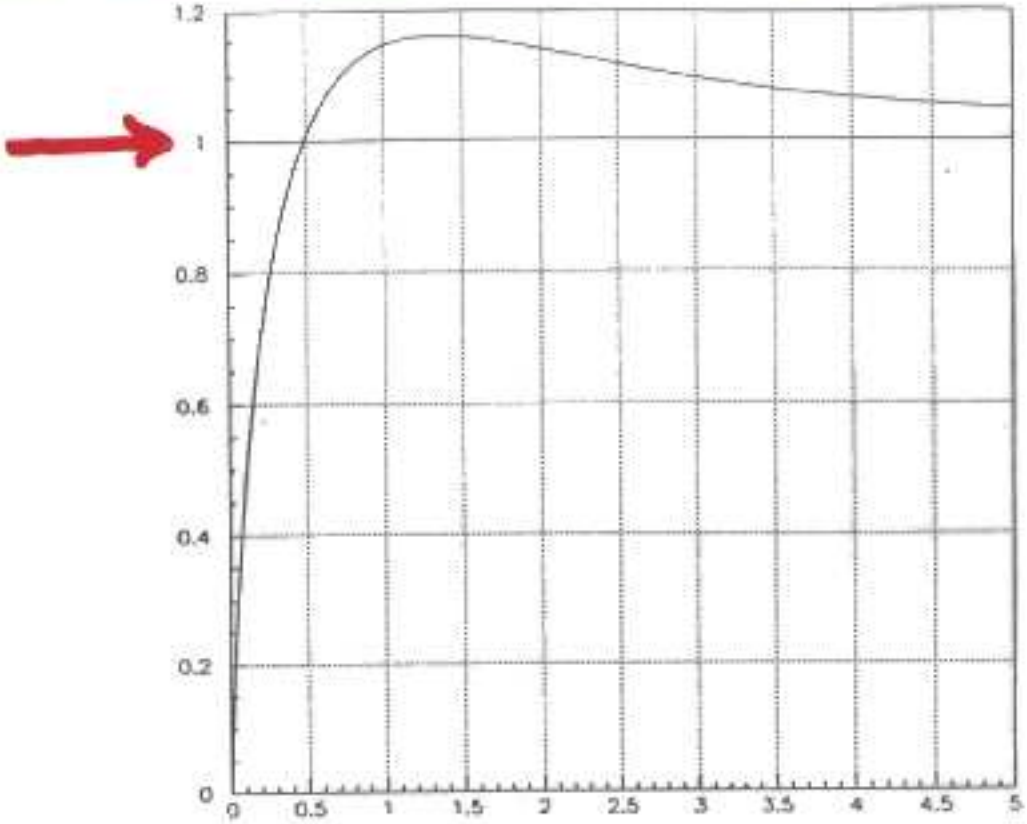
CONCLUSION:

$\mathcal{L}_{max}$ within M.C beam ... NECESSARY

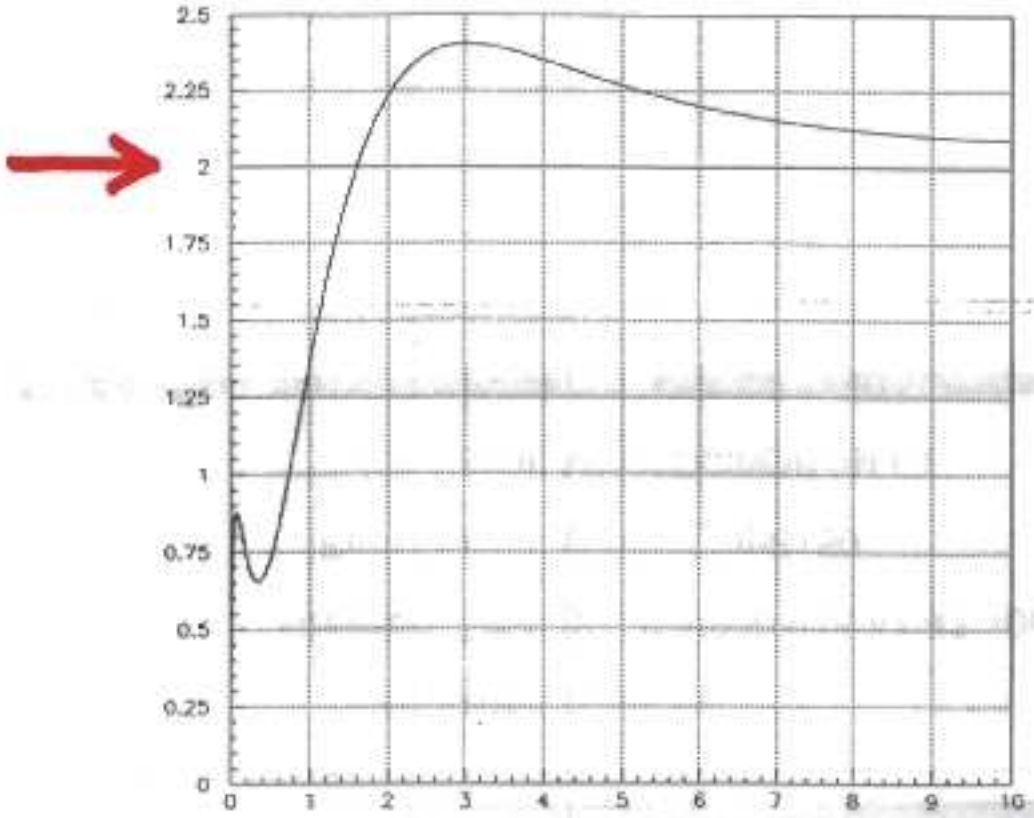
NOT SUFFICIENT

NECESSARY doesn't mean YOU HAVE TO DO IT!

HOW WELL DOES $\ln \mathcal{L}$-ratio approximate χ^2?



$\langle -2 \ln \lambda(\mu; n) \rangle$ vs μ

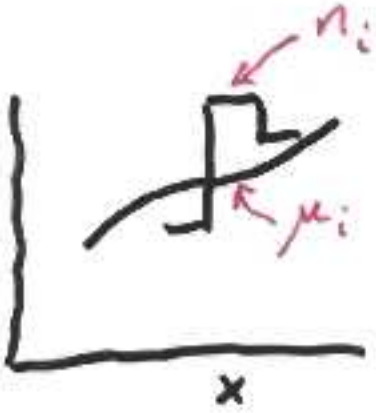


Variance of $-2 \ln \lambda(\mu; n)$ vs μ

JOEL HEINRICH

CDF 5718

BINNED $\mathcal{L}$ & GOODNESS OF FIT



USE $\mathcal{L}$-RATIO

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

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

$$= \prod_i P_{n_i}(n_i)$$

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

$$\xrightarrow{\text{large } \mu_i} -\tfrac{1}{2}\chi^2$$

$\longrightarrow$ Goodness of fit

μ_{best} indep of parameters of fit

$\Rightarrow$ same parameter value from

$\mathcal{L}$ or $\mathcal{L}_{ratio}$

BAKER + COUSINS, NIM A221 (1984) 437

EXTENDED MAXIMUM LIKELIHOOD

Maximum Likelihood uses shape ⇒ params

Extended Max Like uses shape + normalisation

ie. EML uses prob of

1) observing sample size of N events

a 2) given distribution in x.....

⇒ shape parameters + normalisation

Example 1:

Angular distribution

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

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

ML & EML

Maximum Likelihood uses fixed normalisation

Extended Max Like has normalisation as parameter

e.g.1 Decay of resonance

Use M.L for Branching Ratios

Use EML for Partial Decay Rates

e.g.2 Cosmic ray experiment

See 96 protons & 4 heavy nuclei

M.L estimate 96 ± 2% protons 4 ± 2% heavy

EML estimate 96 ± 10 protons 4 ± 2 heavy

a) Max Like

Prob for fixed N = Binomial

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

Prob of forwards

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

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

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

$\Rightarrow$ 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

$$P_b = P_a \times \frac{e^{-\nu} \nu^N}{N!}$$

ν: expected overall rate

Poisson for 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

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

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

"USE $\ln \mathcal{L}$ FOR $\hat{p}$ & $\sigma_{\hat{p}}$

SMEAR IT TO INCORPORATE

SYSTEMATIC UNCERTAINTIES"

SCENARIO:

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

PARAM OF INTEREST — BACKGROUND

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

UNCERTAINTIES MEASURED IN 'SUBSIDIARY' EXPT

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

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

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

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

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

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

7) GETTING χ^2 WRONG

GIOVANNI PUNZI : PHYSTAT 2003

"COMMENTS ON χ^2 FITS WITH VARIABLE RESOLUTION"

SEPARATE SIGNAL FROM BGD

RESOLUTION VARIES EVENT BY EVENT

$\overline{\sigma}$ DIFFERENT FOR SIGNAL + BGD

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

BGD ISOTROPIC OF COSMIC RAYS

i.e. different parts of detector $\Rightarrow$ different σ

2) M (or τ)

Different number of tracks $\Rightarrow$ different σ_M (or σ_τ)

GIOVANNI'S MONTE CARLO FOR

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

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

$f_A = 1/3$

σ_A	σ_B	$\mathcal{L}_x$ f_A	$\mathcal{L}_x$ σ_f	$\mathcal{L}_\checkmark$ f_A	$\mathcal{L}_\checkmark$ σ_f
1.0	1.0	0.336 (3)	0.08		
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) $\mathcal{L}_x$ OK for $p(\sigma_A) = p(\sigma_B)$, but otherwise BIASSED

2) $\mathcal{L}_\checkmark$ gives smaller σ_f than $\mathcal{L}_x$

3) $\mathcal{L}_\checkmark$ unbiassed, but $\mathcal{L}_x$ biassed (enormously !)

Events characterised by x_i & σ_i

A events centred on $x = 0$
B events centred on $x = 1$
} Gaussians with σ_i

$$\mathcal{L}_x(f) = \prod_i \left[f\, G(x_i, 0, \sigma_i) + (1-f)\, G(x_i, 1, \sigma_i) \right]$$

↑ Event by event σ_i

$$\mathcal{L}_\checkmark(f) = \prod_i \left[f\, p(x_i, \sigma_i | A) + (1-f)\, p(x_i, \sigma_i | B) \right]$$

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

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

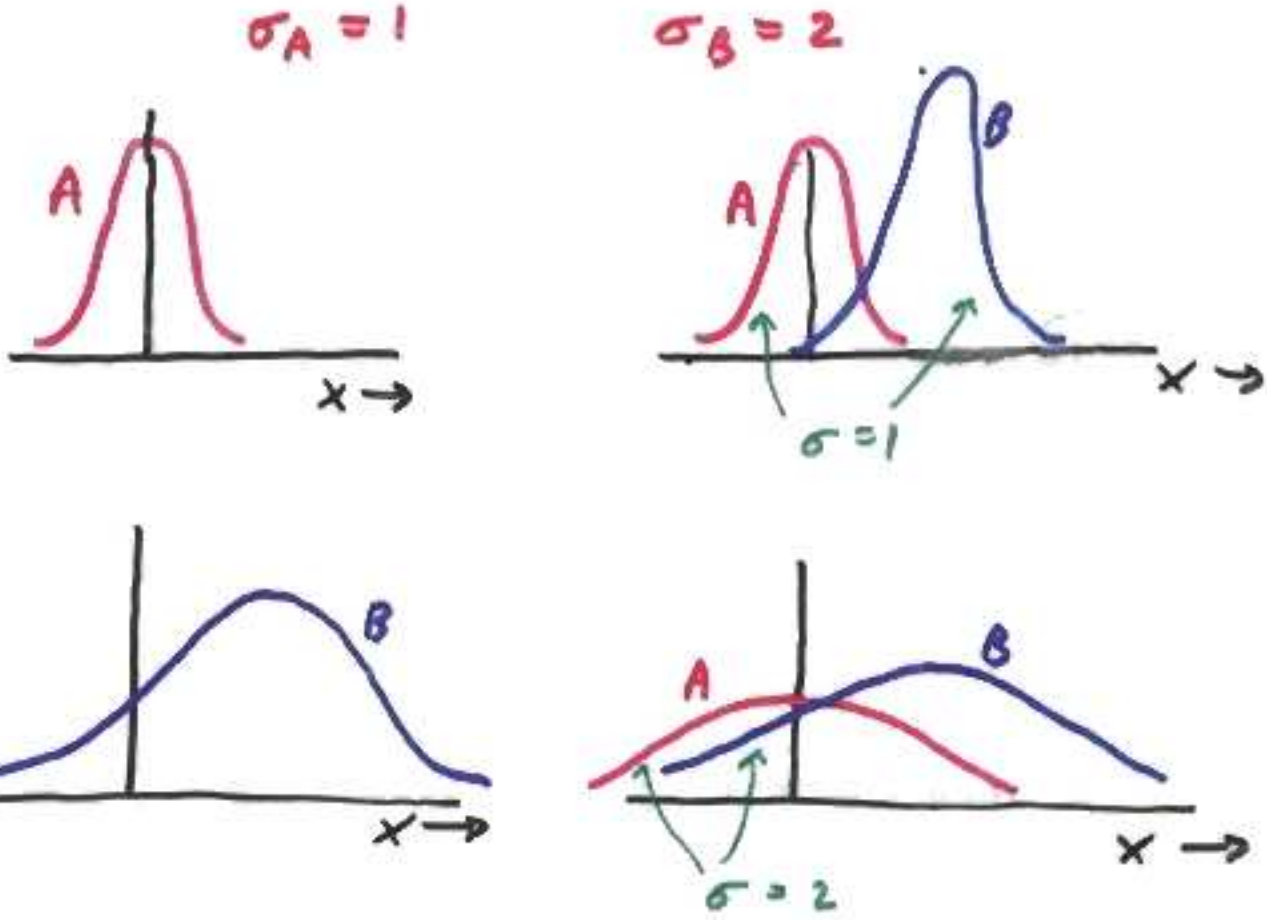
$$= G(x_i, 0, \sigma_i)\, p(\sigma_i | A)$$

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

IF $p(\sigma | A) = p(\sigma | B)$, $\mathcal{L}_\checkmark \equiv \mathcal{L}_x$

BUT NOT OTHERWISE

EXPLANATION OF BIAS



ACTUAL DISTRIBUTION

FITTING FUNCTION
[N_A/N_B VARIABLE, BUT <u>SAME</u> FOR A + B EVENTS]

FIT GIVES <u>UPWARD</u> BIAS FOR N_A/N_B BECAUSE

a) THAT IS MUCH BETTER FOR A EVENTS, +

b) IT DOES NOT HURT TOO MUCH FOR B EVENTS

SOLUTION

INCLUDE $p(\sigma|A)$ & $p(\sigma|B)$ IN FIT

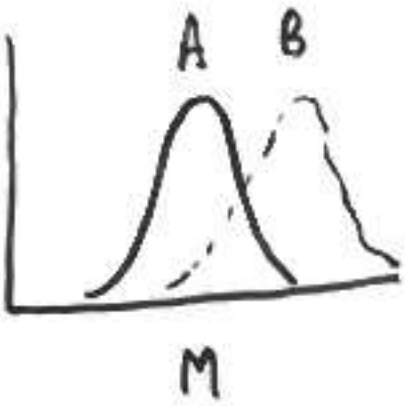
OR

FIT EACH RANGE OF σ_i SEPARATELY, &

ADD $(N_A)_i \rightarrow (N_A)_{tot}$ & $(N_B)_i \rightarrow (N_B)_{tot}$

INCORRECT METHOD USING $\mathcal{L}_x$ USES WEIGHTED AVERAGE OF $(f_A)_i$, ASSUMING INDEPENDENT OF i.

ANOTHER SCENARIO FOR PROBLEM

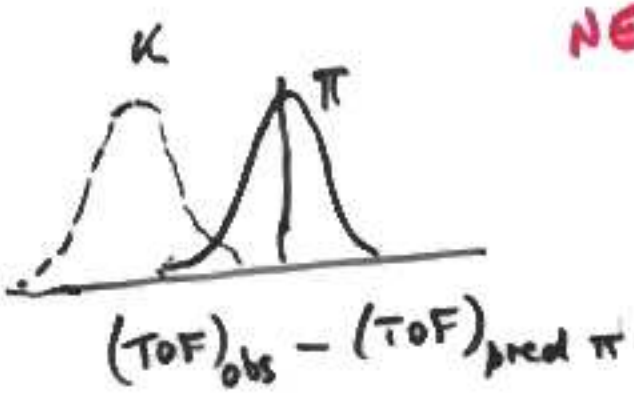


ORIGINALLY:

Positions of peaks = constant

σ_i: variable

$\overline{(\sigma_i)_A} \neq \overline{(\sigma_i)_B}$



NEW SCENARIO : PID

$\sigma_i \sim$ constant

K peak → π peak at large momentum

$\bar{p}_K \neq \bar{p}_\pi$

Common feature : $\frac{\text{SEPARATION}}{\text{ERROR}} \neq$ CONSTANT

WHERE ELSE ???

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

Comments on "Max $\mathcal{L}$" method

1) Uses individual events ⇒ no need to bin

2) Often most efficient method of analysing data

3) Unimportant which variable is used e.g. λ or $\tau = \frac{1}{\lambda}$
   ie. maxima & errors correspond
   ↳ if use $l_{max} - \frac{1}{2}$

4) Worst method computationally
   Needs minimisation
   Normalisation may need recomputation at each step

5) Limits on parameters
   Constraints among params
   } easy to enforce
   but max near boundary can cause trouble

6) Background subtraction difficult

7) Weighted events problematic for errors

8) Hypothesis testing not easy.
   Best is M.C., but not really sufficient.

FNAL STATISTICS LECTURE #3

LEAST SQUARES BEST FIT

STRAIGHT LINE

CORRELATED ERRORS

ERRORS IN x AND y

HYPOTHESIS TESTING BY χ^2

ERRORS OF FIRST + SECOND KIND

KINEMATIC FITTING

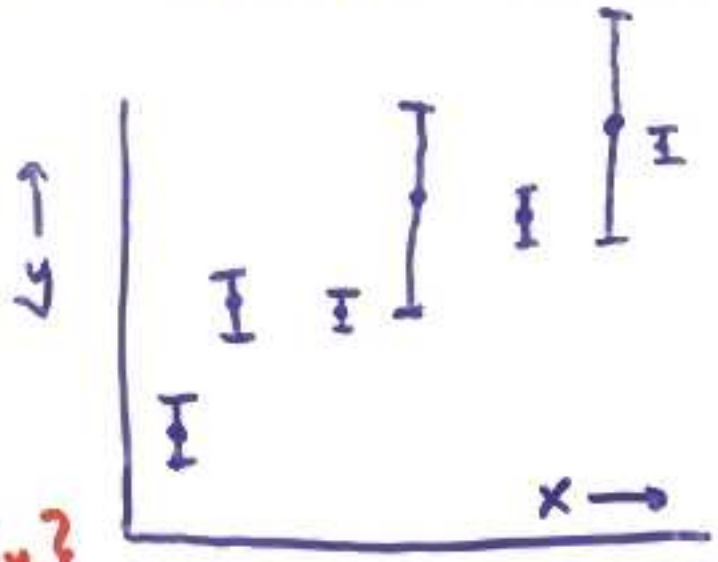
TOY EXAMPLE

THE PARADOX

LOUIS LYONS

AUG 2004

LEAST SQUARES STRAIGHT LINE FITTING



Data
$\{x_i, y_i \pm \delta_{y_i}\}$

Th: $y = a + bx$

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 (& DATA) O.K.

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

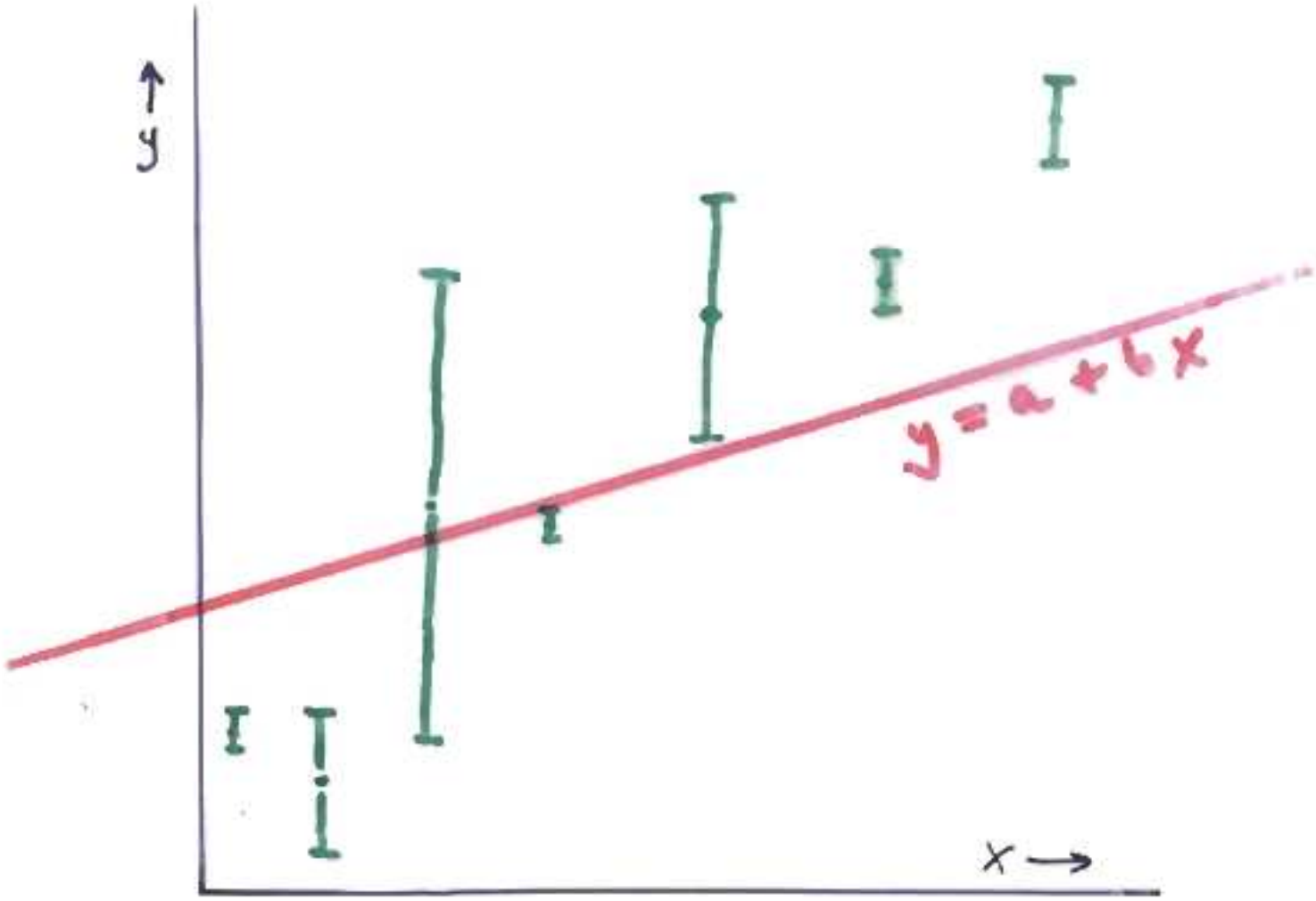
Minimise $S \Rightarrow$ best line

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

* Th Obs σ_{th} σ_{ob} Cont to S

0.01 1 { 0.1 100

1 1



$a + b x_i$ — Vert dev^n

Criterion:

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

σ_i — An error for each pt.

Simple Example of Minimising S

Measurements $a_1 \pm \sigma_1$, $a_2 \pm \sigma_2$, $\vdots$, $a_i \pm \sigma_i$ } Best value $\hat{a} \pm \sigma$

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

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

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

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

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

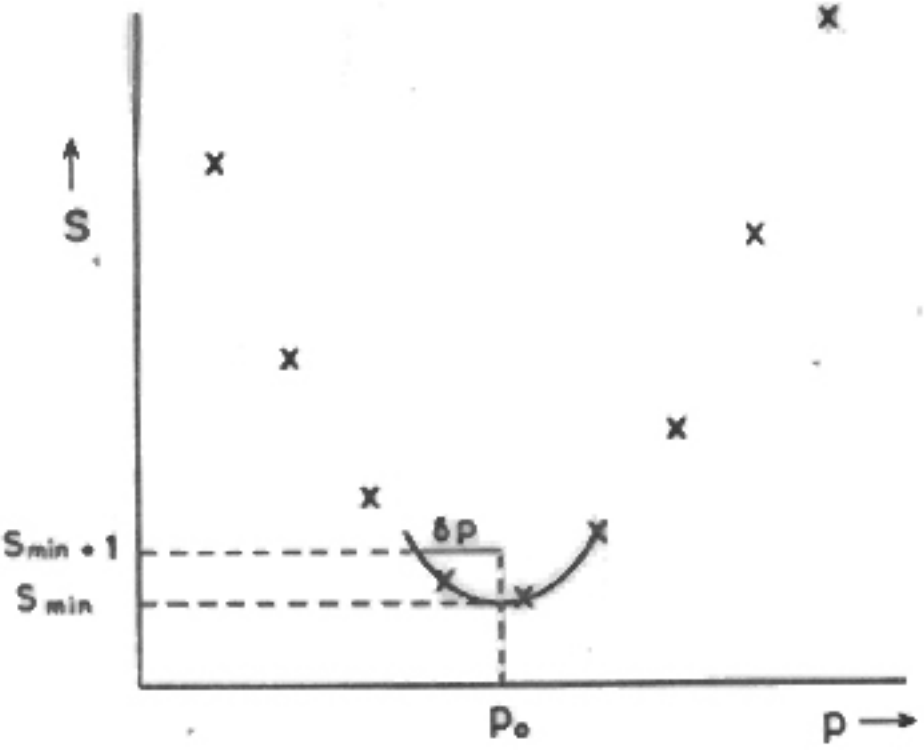
IN PARABOLIC APP EQUIV TO $S \to S_{min} + 1$

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

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

Many params

$\frac{1}{2} \frac{\partial^2 S}{\partial p_i \partial p_j}$ = INVERSE ERROR MATRIX


S
Smin + 1
S min
δp
Po
p

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

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

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

$$\frac{1}{2}\frac{\partial S}{\partial a} = \sum_i \frac{(a + bx_i - y_i)}{\sigma_i^2} = 0$$

$$\frac{1}{2}\frac{\partial S}{\partial b} = \sum_i \frac{(a + bx_i - y_i)x_i}{\sigma_i^2} = 0$$

} 2 SIM. EQNS FOR 2 UNKNOWNS ($\underline{a}$ & $\underline{b}$)

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

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

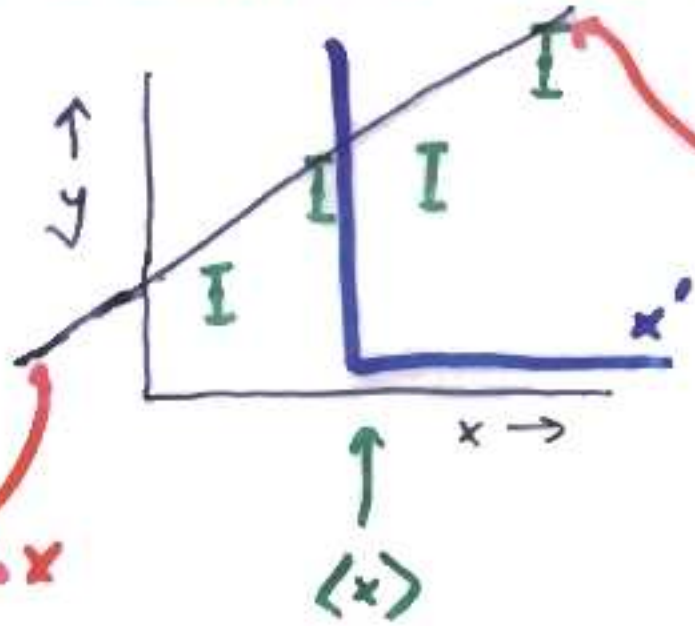
& $\langle f \rangle = [f]/[1]$

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

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

Error on intercept & gradient

First transform so $\langle x \rangle \to 0$



$y = a' + bx'$

$y = a + bx$

$\langle x \rangle$

$x' = x - \langle x \rangle$

$\langle x' \rangle = 0$

Better to use x' because

error on a' & b are UNCORRELATED

[Cf. Errors on a & b CORRELATED]

$$\sigma(a') = 1\Big/\sqrt{\tfrac{1}{2}\frac{\partial^2 S}{\partial a'^2}}$$

$$\sigma(b) = 1\Big/\sqrt{\tfrac{1}{2}\frac{\partial^2 S}{\partial b^2}}$$

$\Leftarrow \operatorname{cov}(a', b) = 0$

$$S = \sum_i \left(\frac{a' + b x_i' - y_i}{\sigma_i}\right)^2 = a'^2[1] + b^2[x'^2] + [y^2] + \text{cross-term (inc } a'b[x'])$$

$$\sigma^2(a') = 1/[1]$$

$$\sigma^2(b) = 1/[x'^2]$$

N.B. Errors depend on σ_i, but NOT on how well data agrees with theory

For error on y at other x', use $y = a' + bx'$

$$\Rightarrow \sigma^2(y) = \sigma^2(a') + x'^2\,\sigma^2(b)$$

For $x' = -\langle x \rangle$ [i.e. $x = 0$]

$$\sigma^2(a) = \sigma^2(a') + \langle x \rangle^2\,\sigma^2(b)$$

BUT $\sigma(a)$ & $\sigma(b)$ CORRELATED

SPECIAL CASE : All σ_i same

$$\sigma^2(a') = \sigma^2/n$$

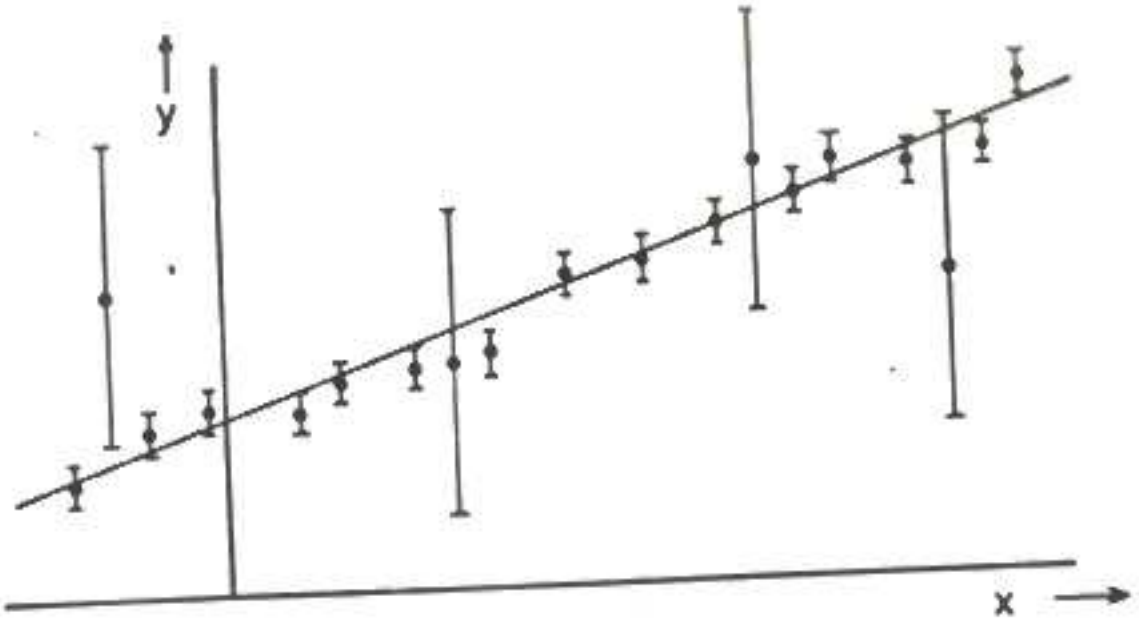
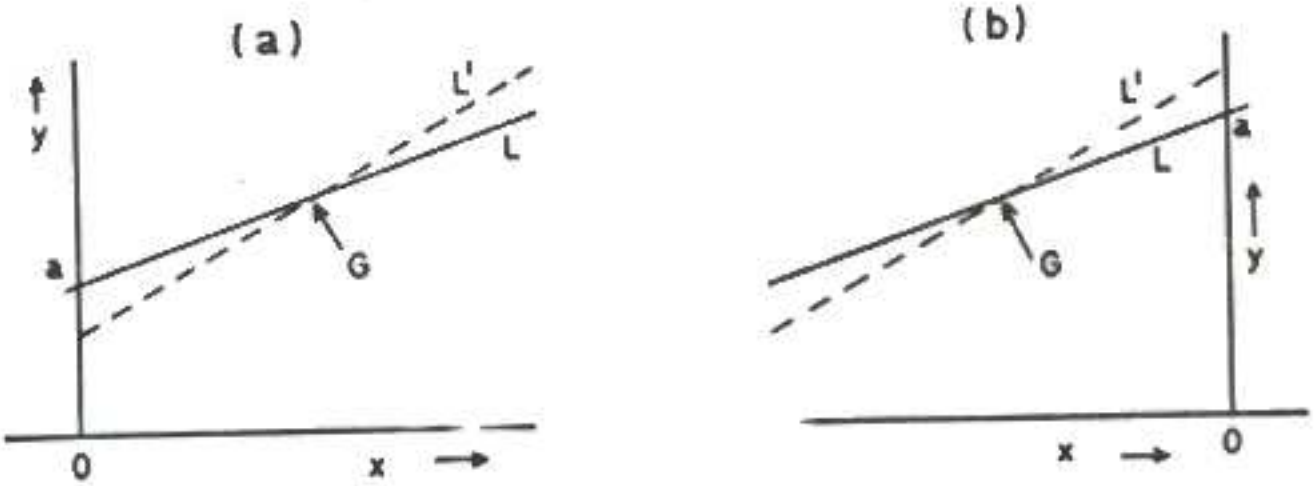


Fig. 2.3

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



$\langle x \rangle$ pos

Fig. 2.4

$\langle x \rangle$ neg

IF NO ERRORS δy_i (!)

ASSUME ALL ERRORS EQUAL
(or similar)

σ CANCELS FROM a & b

e.g. $b = \dfrac{[1][xy] - [x][y]}{[1][x^2] - [x]^2}$

NEED σ for errors on a' & b

$$S = \frac{1}{\sigma^2}\sum (a + bx_i - y_i)^2 = \nu$$

$\Rightarrow \sigma$

$\Rightarrow \sigma(a')$ & $\sigma(b)$

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

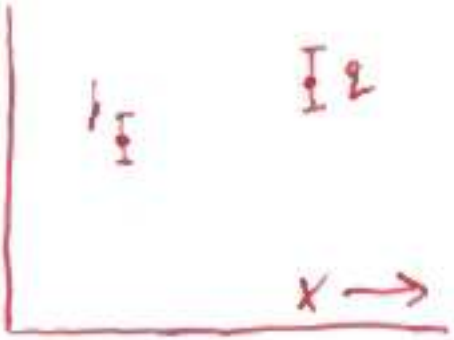
SUMMARY OF STRAIGHT LINE FIT

1) PLOT DATA
   a) BAD POINTS
   b) a AND b . & $\sigma(a')$, $\sigma(b)$
2) a AND b FROM FORMULAE *
3) ERRORS on a' AND b *
4) CF 2) and 3) WITH 1)
5) DETERMINE S_{MIN} (using a & b) *
6) $\nu = n - p$ *
7) Look up χ^2 tables *
8) IF PROBABILITY TOO SMALL, IGNORE RESULTS
8a) IF PROBABILITY is "A BIT" SMALL, SCALE ERRORS?

* COMPUTER PROGRAMME

MEASUREMENTS WITH CORRELATED ERRORS

e.g. systematics ?



Start with 2 uncorrelated measurements

$$S = \frac{(p - p_{pr})^2}{\sigma_p^2} + \frac{(q - q_{pr})^2}{\sigma_q^2} \quad *$$

Introduce correlations by

$$p = r\cos\theta - s\sin\theta$$
$$q = r\sin\theta + s\cos\theta$$

} NOT ROTN in x-y SPACE

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

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

Inv. err matrix element $= H_{11}(r - r_{pr})^2 + H_{22}(s - s_{pr})^2 + 2H_{12}(r - r_{pr})(s - s_{pr})$

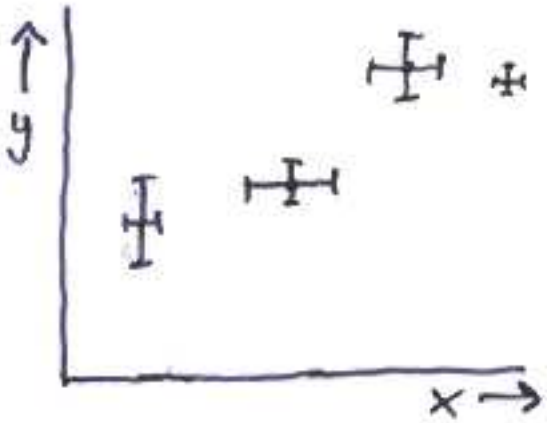
show $H^{-1} = \begin{pmatrix} \sigma_r^2 & \operatorname{cov} \\ \operatorname{cov} & \sigma_s^2 \end{pmatrix}$ ← Error matrix

Reduces to standard formula in absence of corrlns

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

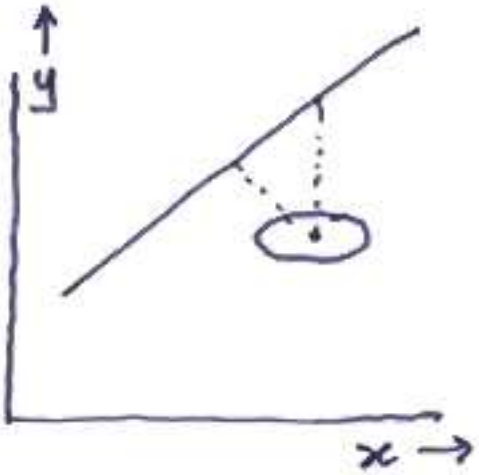
show $\Delta_j = (\text{observed} - \text{pred.})_j$

STRAIGHT LINE : ERRORS on x AND y



For simplicity,

assume x, y errors uncorrelated



Previously, contribution to S was $\left(\frac{y_i - y_i(fit)}{\sigma_i}\right)^2$

Now replace by

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

ie. 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 $n \gtrsim 3$

3) No problem with correlated errors

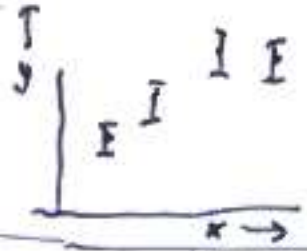
4) Can calculate χ^2 "on line" (ie. single pass through data)

$$\sum \frac{(y_i - a - bx)^2}{\sigma^2} = [y_i^2] - b[x_i y_i] - a[y_i]$$

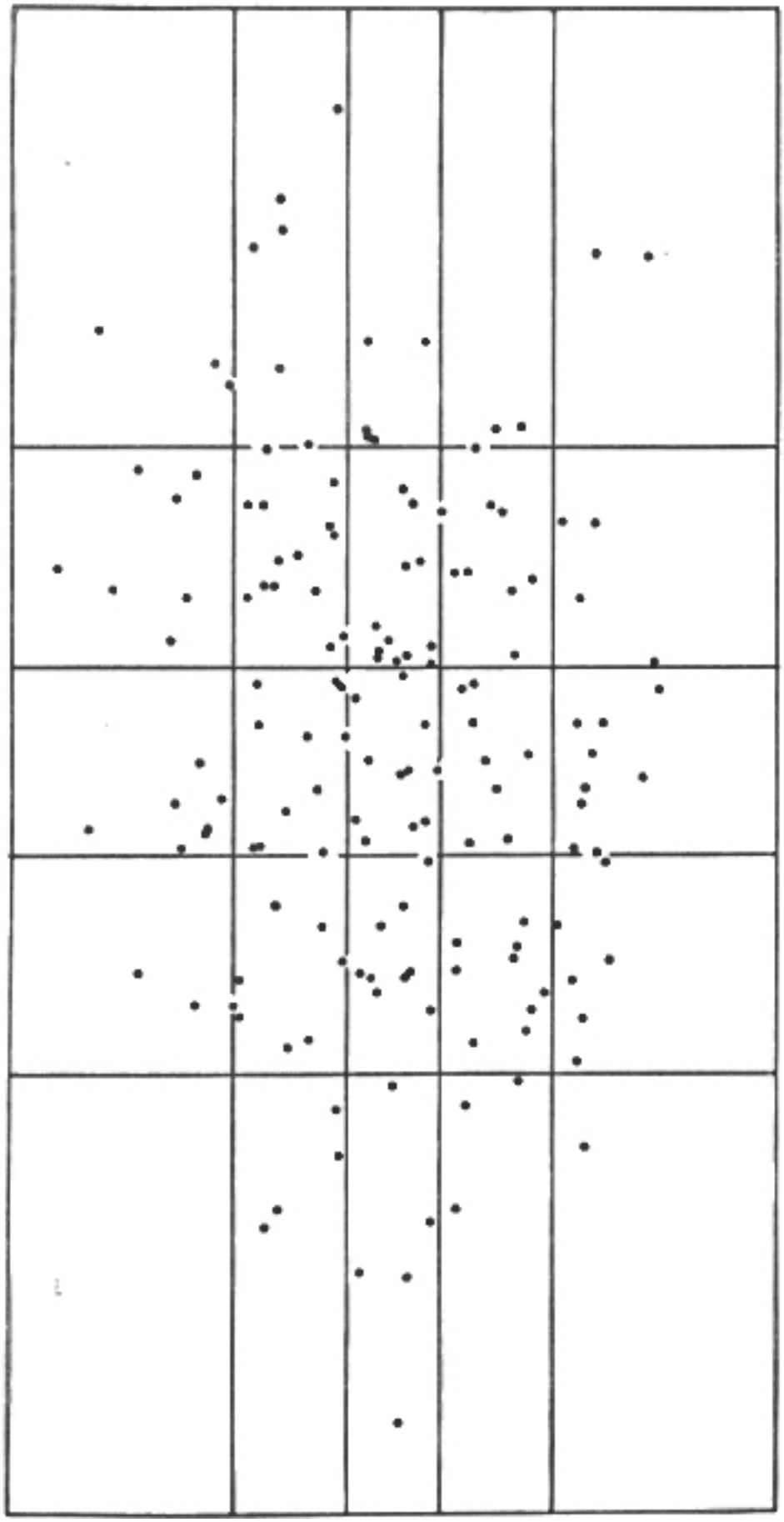
5) For theory linear in parameters,

soln can be found analytically

6) Hypothesis Testing ✡ ✡ ✡



	Individual events (e.g in $\cos\theta$)	$y_i \pm \sigma_i$ v x_i (e.g. stats)
1) Binning first	✓	×
4) χ^2 on line	First histogram	✓

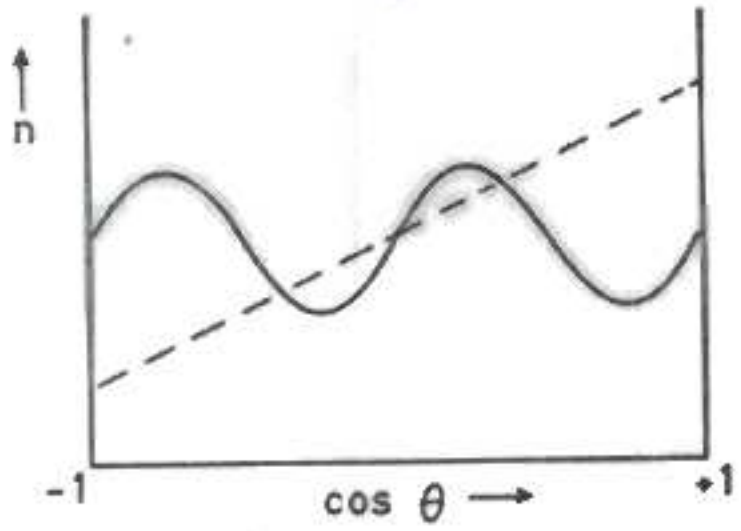


	Nom.	M. L.	L. S.
Easy?	Yes, if...	Nom, mosm. messy	Minimisation
Efficient?	Not very	Usually best	Sometimes ≡ M.L
Input	Separate evnt.	Separate ev.	Histogram
Goodness of Fit	Messy	V. difficult	Easy
Constraints	No	Easy	Can be done
n-dimensions	Easy, if...	Nom, mosm messier	Needs v. many events
Weighted ev.	Easy	Errors diff.	Easy
Bgd sub	Easy	Troublesome	Easy
Error est.	Observed spread OR Analytic	$\left(-\frac{\partial^2 l}{\partial p_i \partial p_j}\right)^{-\frac{1}{2}}$	$\left(\frac{1}{2}\frac{\partial^2 S}{\partial p_i \partial p_j}\right)^{-1}$
Main +	EASY	BEST FEW EVENTS	HYP. TEST.

HYPOTHESIS TESTING

by PARAMETER TESTING

$1 + \frac{b}{a}\cos^2\theta$ Is $\frac{b}{a} = 0$?



"DISTRIBUTION TESTING" IS BETTER

HYPOTHESIS TESTING

χ^2 TEST

1) CONSTRUCT S, + MINIMISE W.R.T. FREE PARAMETERS

2) DETERMINE ν = No. OF DEGREES OF FREEDOM

$$\nu = n - p$$

n = No OF DATA POINTS

p = No OF FREE PARAMS

3) LOOK UP PROB THAT, FOR ν DEG OF FREEDOM, $\chi^2 \geq S_{min}$

[ASSUMES y_i ARE GAUSSIAN DISTRIBUTED WITH MEAN y_i^{th} AND VARIANCE σ_i^2]

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

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

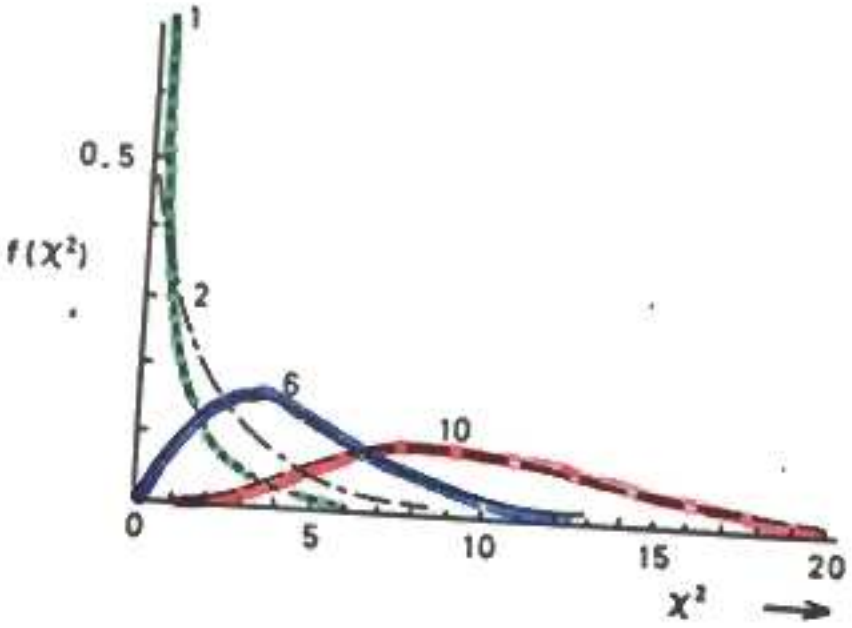
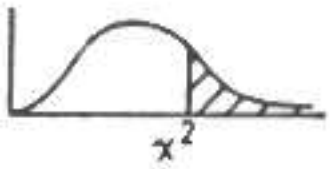


Fig. 2.6

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

is LARGE

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



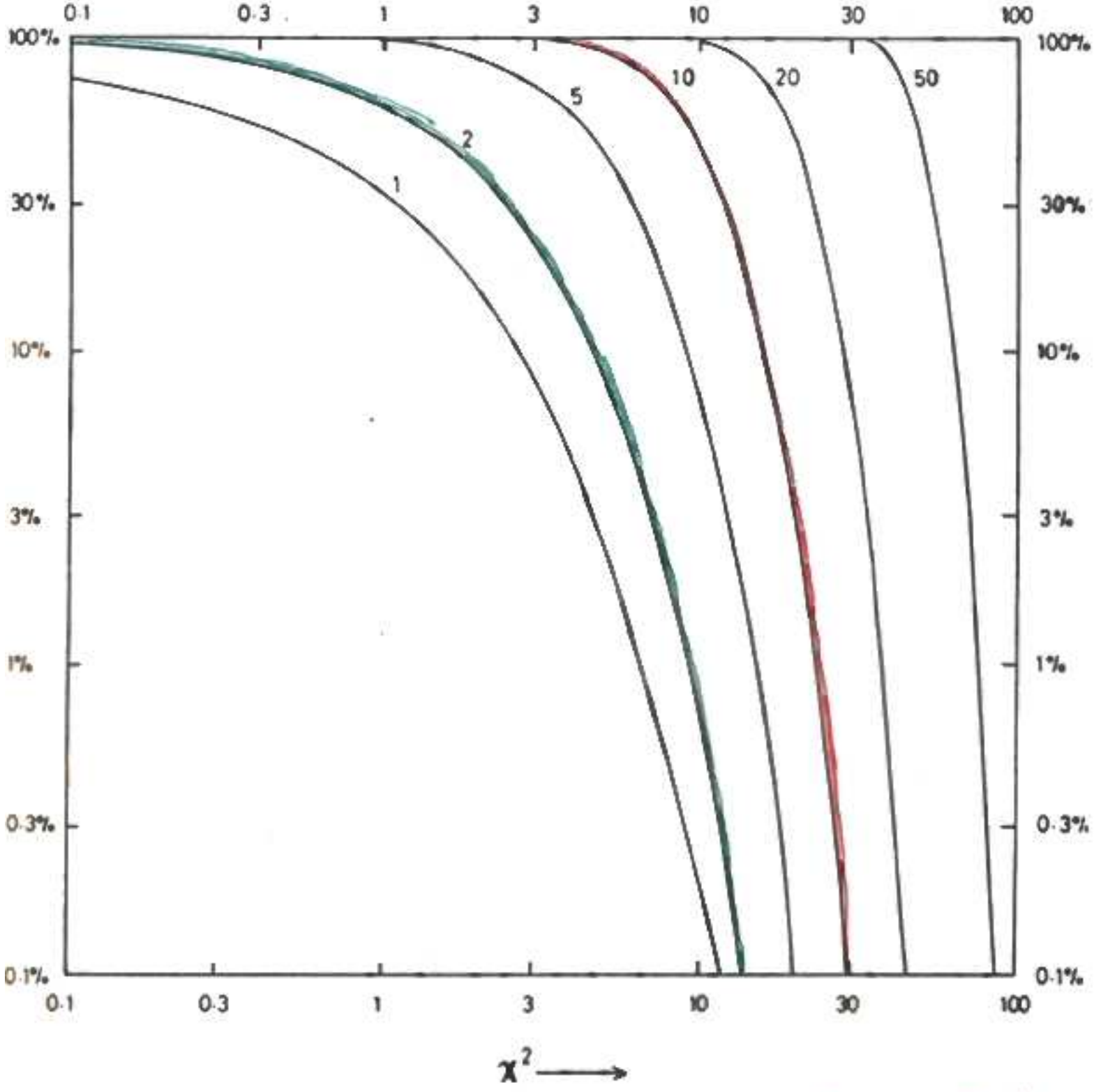


Fig. 2.7

CF: Area in tails of Gaussian

Goodness of Fit

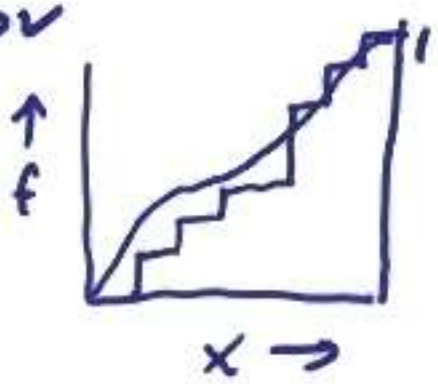
χ^2 : Very general

Needs binning

Not sensitive to sign of devn.

Run test

Kolmogorov - Smirnov



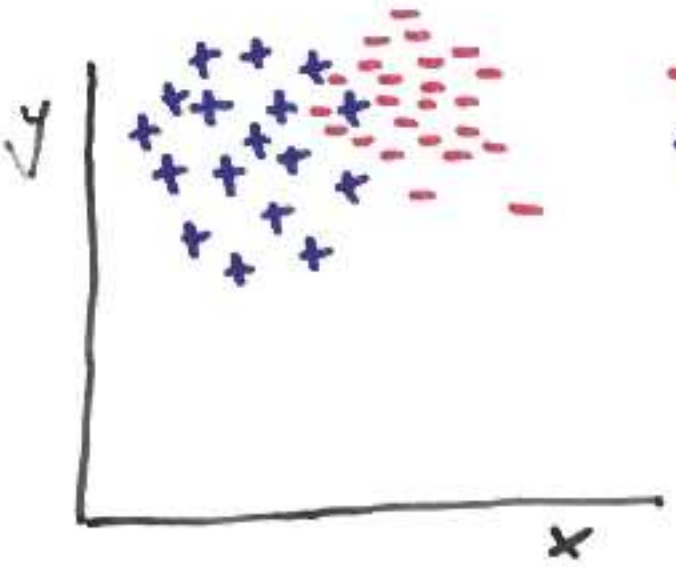
etc

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

Maria Grazia Pia's group in Genoa

ENERGY TEST FOR GOODNESS OF FIT

Aslan + Zech



$$\text{"Energy"} = \sum_{i,j} q_i q_j f(r_{ij})$$

$$f = \frac{1}{r+\epsilon}$$

$$\text{or } -\ln(r+\epsilon)$$

N.B. ϵ, choice of f

Scaling of $x, y, \ldots$

Need M.C.

WRONG DECISIONS

ERROR OF FIRST KIND

Reject H when it is true

Should happen x % of time

ERROR OF SECOND KIND

Accept H when something else is true

How often depends on .....

i) How similar other hypotheses are

e.g. $H = \pi$

Alternatives = e μ K p ...

ii) Relative frequencies

e.g. 10^{-4} 10^{-4} 10% 10%

Aim for maximum effic ← small error 1st kind

maximum purity ← small error 2nd kind

As χ^2 cut increases, effic ↑ purity ↓

Choose compromise

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

1) Result of experiment

e.g. Is spin of resonance = 2?

GET ANSWER WRONG

Where to set χ^2 cut?

Large cut : "Never" reject anything

Small cut : Reject when correct

Depends on nature of hypothesis

e.g. Does our result agree with that of exptl E...?

OR Is our data consistent with Special Relativity?

2) Event selector

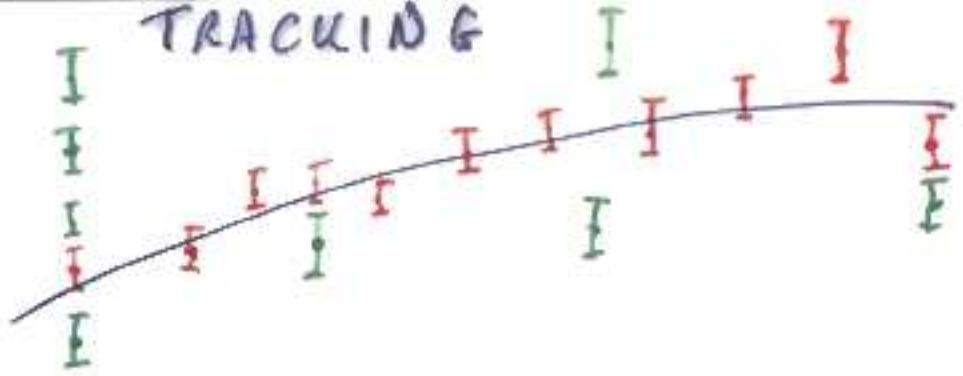
e.g. Does this event contain Z^0?

Error of 1st kind : Loss of effic

Error of 2nd kind : Bgd

Usually easier to allow for 1 than 2.

3) Pattern recognition



Hypothesis Testing = Pattern Recognition

= Find hits that make track

Parameter Determination = Estimate track parameters

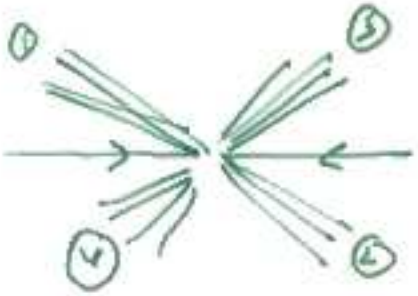
(+ error matrix)

KINEMATIC FITTING

Test whether observed event consistent with specified reaction

e.g. 

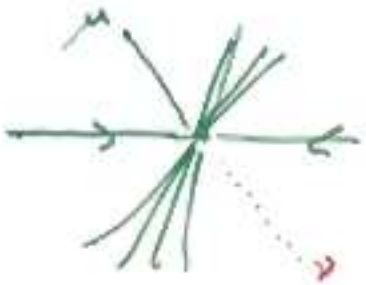
$\bar{p}p \to \bar{p}p\pi^+\pi^-$?



$e^+e^- \to W^+W^-$

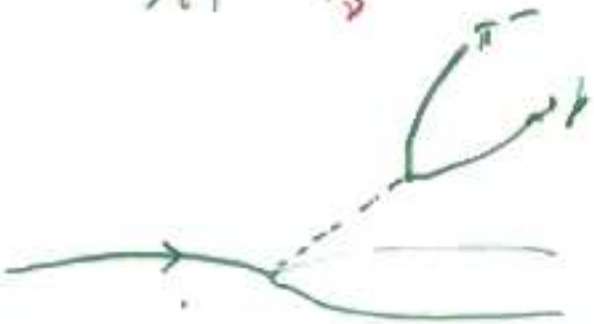
$\hookrightarrow j_1 j_2$ $\quad j_3 j_4$

M_W, jet pairings

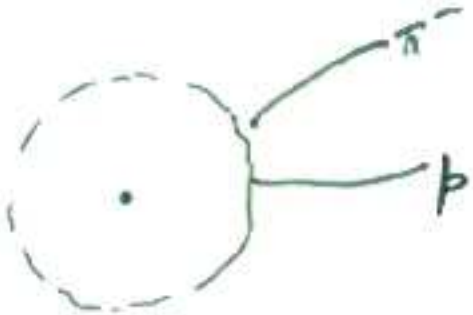


$e^+e^- \to W^+W^-$

$\hookrightarrow j_1 j_2$ $\quad \mu\nu$



$\Lambda \to p\pi^-$ from prodn vertex



$p + \pi^-$ intersect

& $\Lambda \to p\pi^-$ from prodn vert.

WHY DO IT ?

1) CHECK WHETHER EVENT CONSISTENT WITH HYPOTHESIS [HYPOTHESIS TESTING]

2) CAN CALCULATE MISSING VARIABLES [PARAM DETN.]

3) GOOD TO HAVE TRACKS CONSERVING E-P [P.D]

4) IMPROVES ERRORS [P.D]

WHY DO IT ?

1) CHECK WHETHER EVENT CONSISTENT WITH HYPOTHESIS [HYPOTHESIS TESTING]

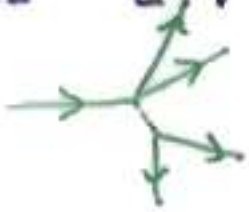
Use S_{min} & No of ~~constraints~~ degrees of freedom

2) CAN CALCULATE MISSING VARIABLES [PARAM DETN.]

e.g |P| for straight / short track / incoming ν

3 momentum of π, ν, ...

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 $p\bar{p} \to p\bar{p}\pi^+\pi^-$ ✡

4 momenta of each track

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

Then test hypothesis:

Observed event = example of reaction ✡

Tested by:

Observed tracks should conserve 4-p

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

ie. $S_{min} = \sum_{\substack{4\ tracks \\ \times 4\ E\text{-}p}} \left(\frac{V_i^{fitted} - V_i^{meas}}{\sigma_i} \right)^2$ ← if uncorr.

Otherwise use Inv. Err. Matrix

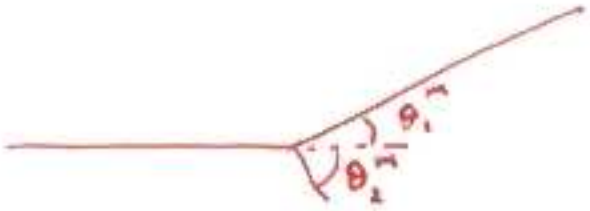
where V_i^{fitted} conserve 4-momenta

i.e. Minimisation subject to constraint

(involves Lagrange multipliers)

TOY EXAMPLE OF FIT

$\bar{p}p \to \bar{p}p$



4 constraints:

1) Coplanar
2) p_1 at θ_1
3) p_2 at θ_2
4) θ_1 at θ_2 $\Longleftarrow$ 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 $S(\theta_1, \theta_2) = \dfrac{(\theta_1 - \theta_1^m)^2}{\sigma^2} + \dfrac{(\theta_2 - \theta_2^m)^2}{\sigma^2}$

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

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

$\Rightarrow$ 3 eqns for θ_1 θ_2 λ

Eqns simple to solve because

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

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

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

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

i.e. KINEMATIC FIT $\Rightarrow$

REDUCED ERRORS

$$\lambda = \frac{\theta_1^m + \theta_2^m - \pi/2}{\sigma^2}$$

PARADOX

Histogram with 100 bins

Fit with one parameter

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

For our data, $S_{min}(b_0) = 85$

Is b_1 acceptable if $S(b_1) = 110$?

1) YES

Very acceptable χ^2 probability

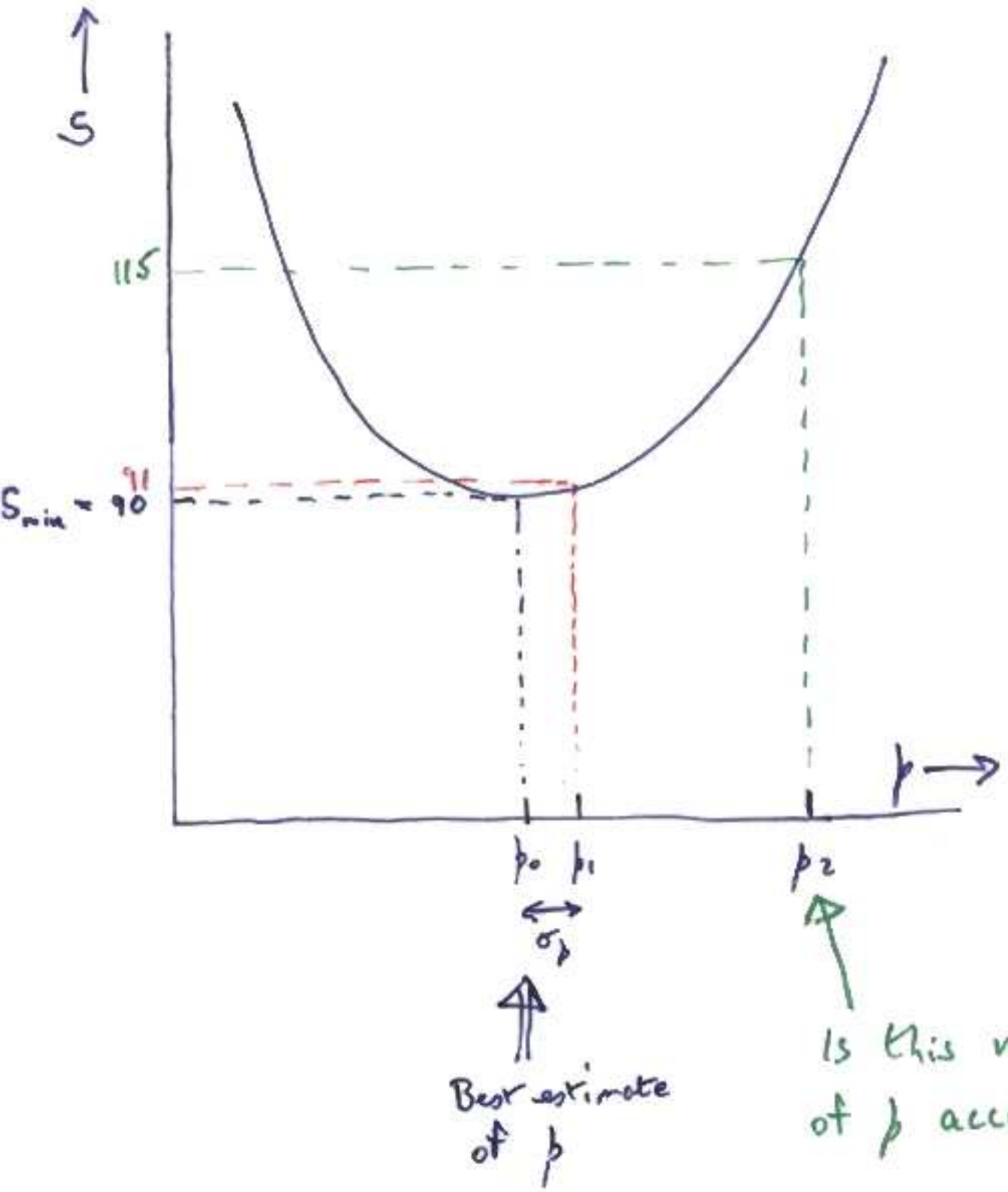
2) NO

σ_b from $S(b_0 \pm \sigma_b) = S_{min} + 1$

$= 86$

But $S(b_1) - S_{min}(b_0) = 25$

$\therefore$ 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

SELECTING BETWEEN TWO HYPOTHESES

Louis Lyons

OUNP-99-12

MATHEMATICAL FORMULATION

$$S(x) = \sum \frac{(x_i - x)^2}{\sigma^2} \equiv \sum \frac{(x_i - \bar{x})^2}{\sigma^2} + N\frac{(\bar{x} - x)^2}{\sigma^2}$$

SCATTER OF POINTS WRT THEIR MEAN.

INDEP OF x

THIS IS TERM WHICH HAS EXPECTED VALUE

$(N-1) \pm \sqrt{2(N-1)}$

χ^2_{N-1}

HOW WELL x AGREES WITH $\bar{x}$

VARIES WITH x

BEST VALUE IS $x = \bar{x}$

INCREASES BY 1 FOR $x = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$

χ^2_1

CONCLUSION FOR THIS CASE

COMPARING $H_1 : p = p_1$

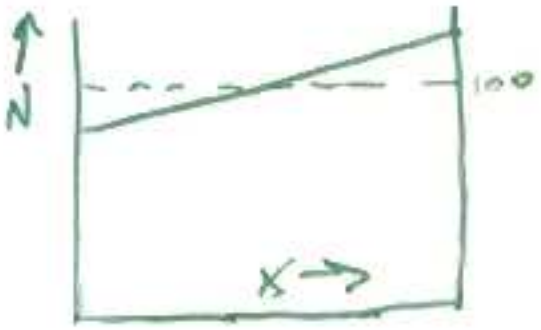
& $H_2 : p = p_2$

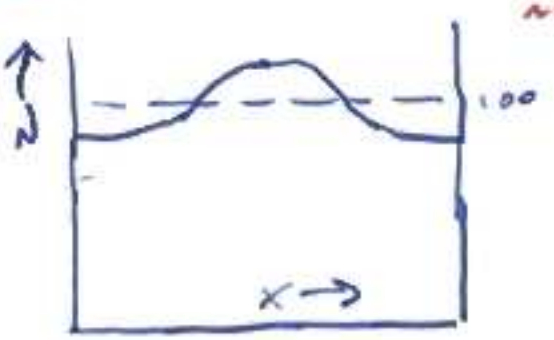
DECISION DEPENDS ON $\Delta\chi^2$

ANOTHER EXAMPLE

$x = -1 \rightarrow +1$

100 bins
~100 events/bin
NO FREE PARAMS





H1 : $1 + ax$
$a = 0.05$

H2 : $1 + b\cos(\pi x)$
$b = 0.05$

Generate events according to H1 (+ stat fluctn)

Try fitting according to H1 or to H2

χ^2_1 χ^2_2

Look at dist of χ^2_1 As expected for NDF=100

χ^2_2 Bit bigger. Many* "satisfactory"

$\chi^2_2 - \chi^2_1$ Decision based on $\Delta\chi^2$ has much better power

Repeat for events generated according to H2

Look at dist of χ^2_1

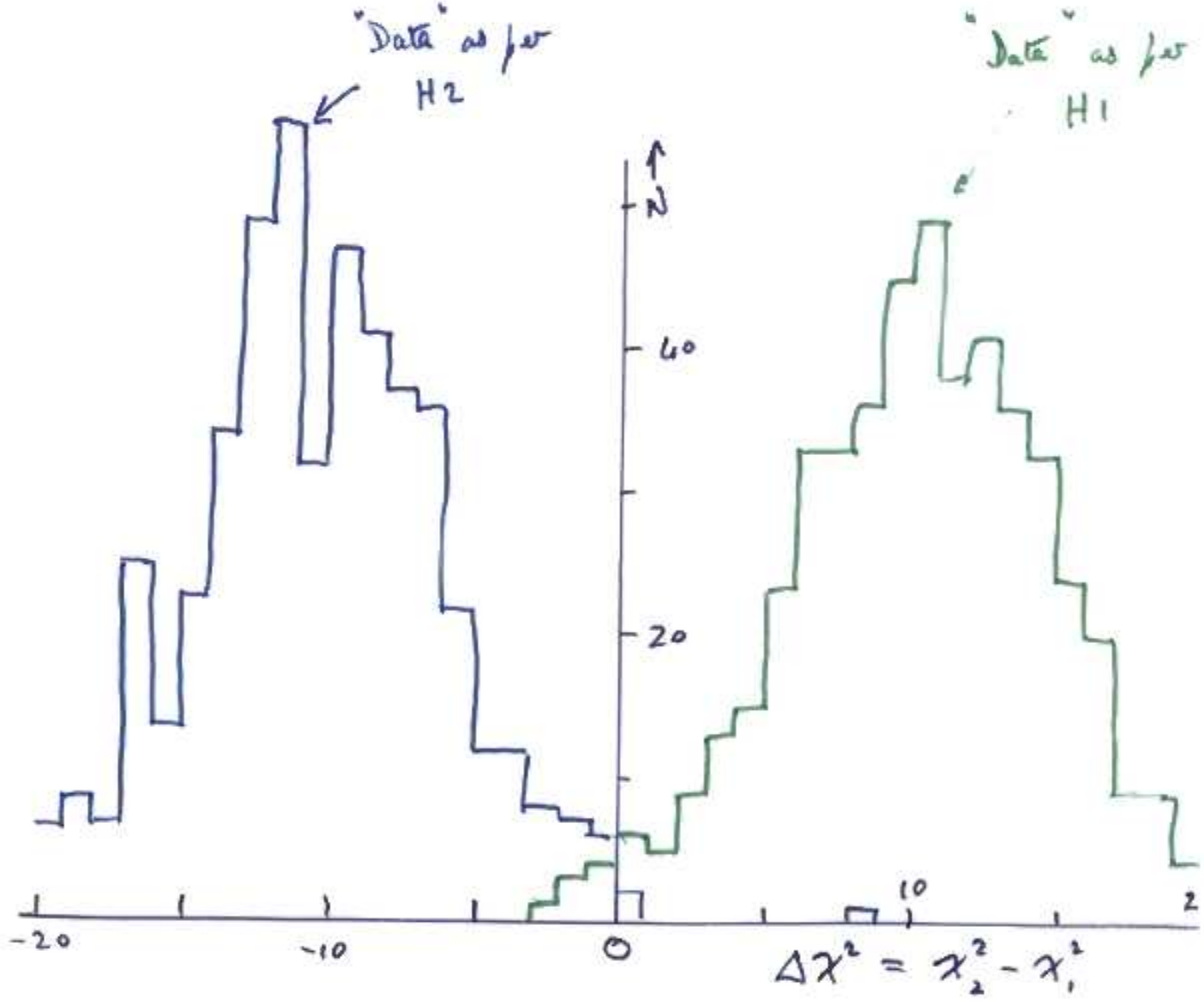
χ^2_2

$\chi^2_2 - \chi^2_1$

* 69% have $\chi^2_2 < 130$

DISTINGUISHING 2 HYPOTHESES ON BASIS OF $\Delta\chi^2$

(500 SIMULATIONS)



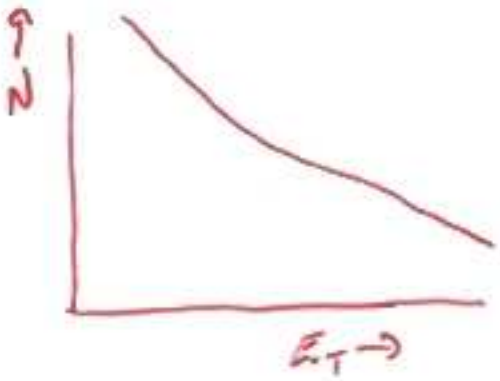
$H2 = 1 + 0.05\cos(\pi x)$

$H1 = 1 + 0.05x$

BAYESIAN

POSSIBLE APPLICATIONS

1) JET E_T DISTRIBUTION AT COLLIDER



FIT DISTRIBUTION BY ALTERNATIVE HYPOTHESES
(DIFFERENT STRUCTURE FNS.)

LOOK AT χ^2_1 FOR STR FN 1

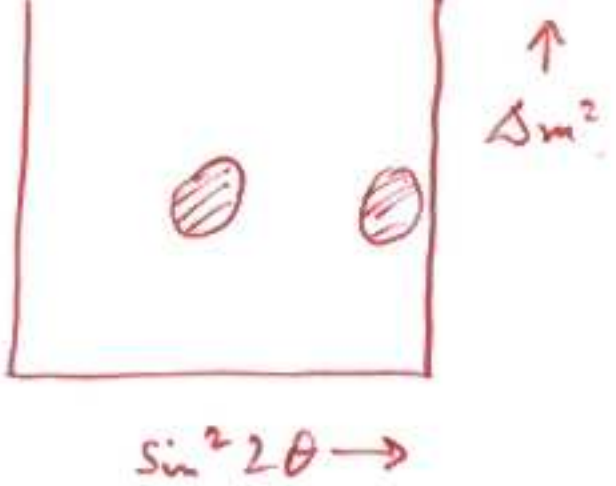
χ^2_2 FOR STR FN 2

DECIDE BETWEEN STR. FNS. ON BASIS OF

$\Delta\chi^2$

EVEN IF LARGER χ^2 GIVES O.K. PROB

2) ν OSCILLATIONS



LARGE ANGLE
SMALL ANGLE
} SOLNS FOR SOLAR NEUTRINOS

(FROM OVERALL RATES FROM DIFFERENT DETECTORS)

LOOK AT DISTRIBUTION OF EVENTS IN SOME VARIABLE

e.g. N

or Night/Day time

or Earth → Sun distance

(Different solns give slightly different shapes)

Assume distribution has χ_1^2 for soln 1
χ_2^2 ---- 2

Choose between solns on basis of $\Delta\chi^2$,

NEED MORE TO DECIDE even if larger χ^2 has OK probability

FNAL STATISTICS LECTURE #4

LOUIS LYONS

BAYES ✓ FREQUENTISM

BAYES versus FREQUENTISM

The Return of an Old Controversy

- The ideologies, with examples
- Upper limits
- Systematics

Louis Lyons, Oxford University
and CERN

It is possible to spend a lifetime analysing data without realising that there are two very different approaches to statistics:

Bayesianism and Frequentism.

How can textbooks not even mention

Bayes/ Frequentism?

For simplest case $(m \pm \sigma) \leftarrow Gaussian$

with no constraint on $m(true)$ then

$$m - k\sigma < m(true) < m + k\sigma$$

at some probability, for both Bayes and Frequentist

(but different interpretations)

See Bob Cousins "Why isn't every physicist a Bayesian?" Amer Jrnl Phys 63(1995)398

We need to make a statement about

Parameters, Given Data

The basic difference between the two:

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

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

PROBABILITY

MATHEMATICAL

Formal

Based on Axioms

FREQUENTIST

Ratio of frequencies as $n \rightarrow$ 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"

Bayesian versus Classical

Bayesian

$P(A \text{ and } B) = P(A;B) \times P(B) = P(B;A) \times P(A)$

e.g. A = event contains t quark

B = event contains W boson

or A = you are in CERN

B = you are at Workshop

Completely uncontroversial, provided….

$P(A;B) = P(B;A) \times P(A) / P(B)$

Bayesian $$P(A;B) = \frac{P(B;A) \times P(A)}{P(B)}$$ Bayes Theorem

$$P(hyothesis; data) \propto P(data; hypothesis) \times P(hypothesis)$$

posterior ↑ likelihood ↑ prior ↑

Problems: P(hyp..) true or false

"Degree of belief"

Prior What functional form?

Coverage

Goodness of fit

P(hypothesis……) True or False

"Degree of Belief"

credible interval

Prior: What functional form?

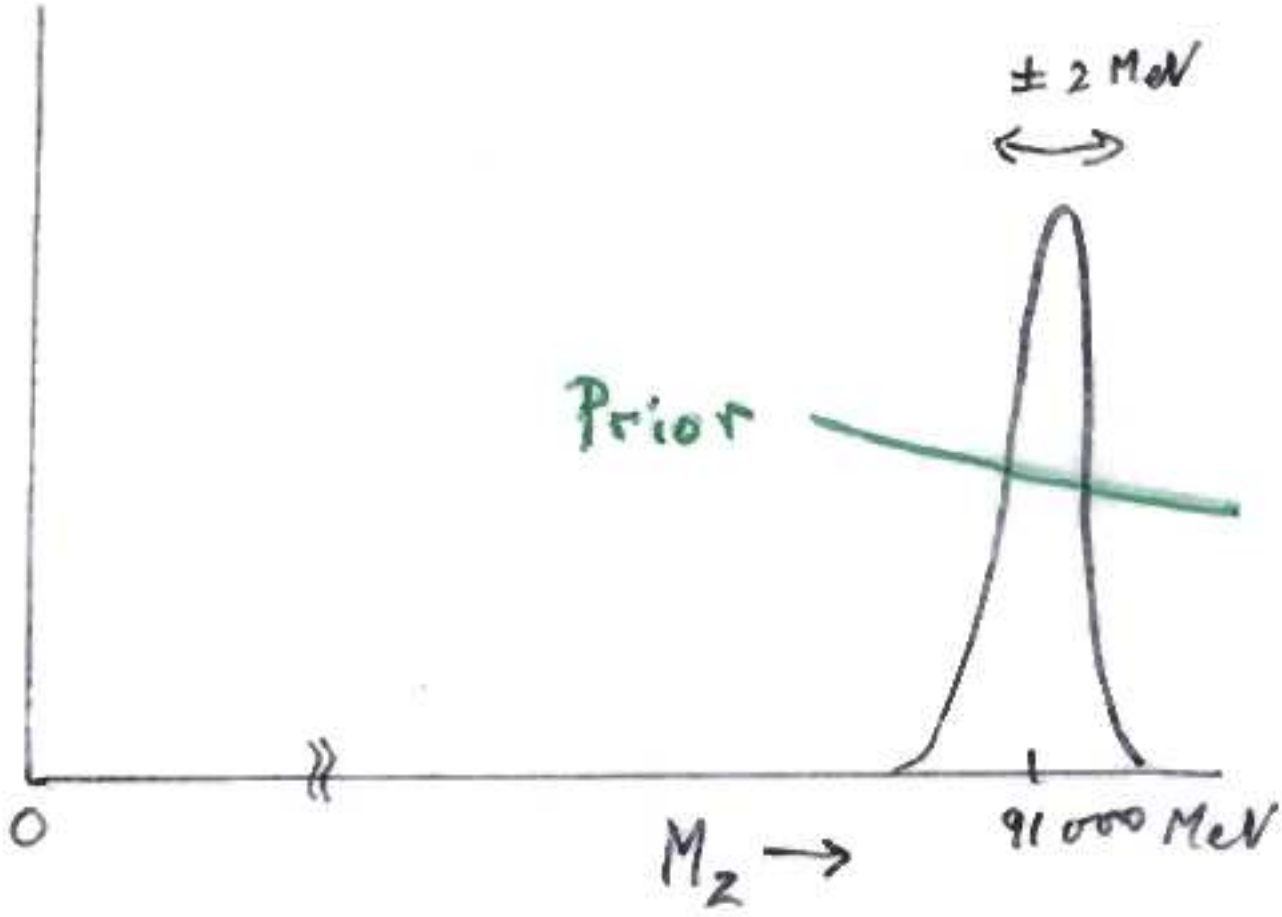
Uninformative prior:

flat? In which variable? e.g. $m, m^2, \ln m, \ldots.?$

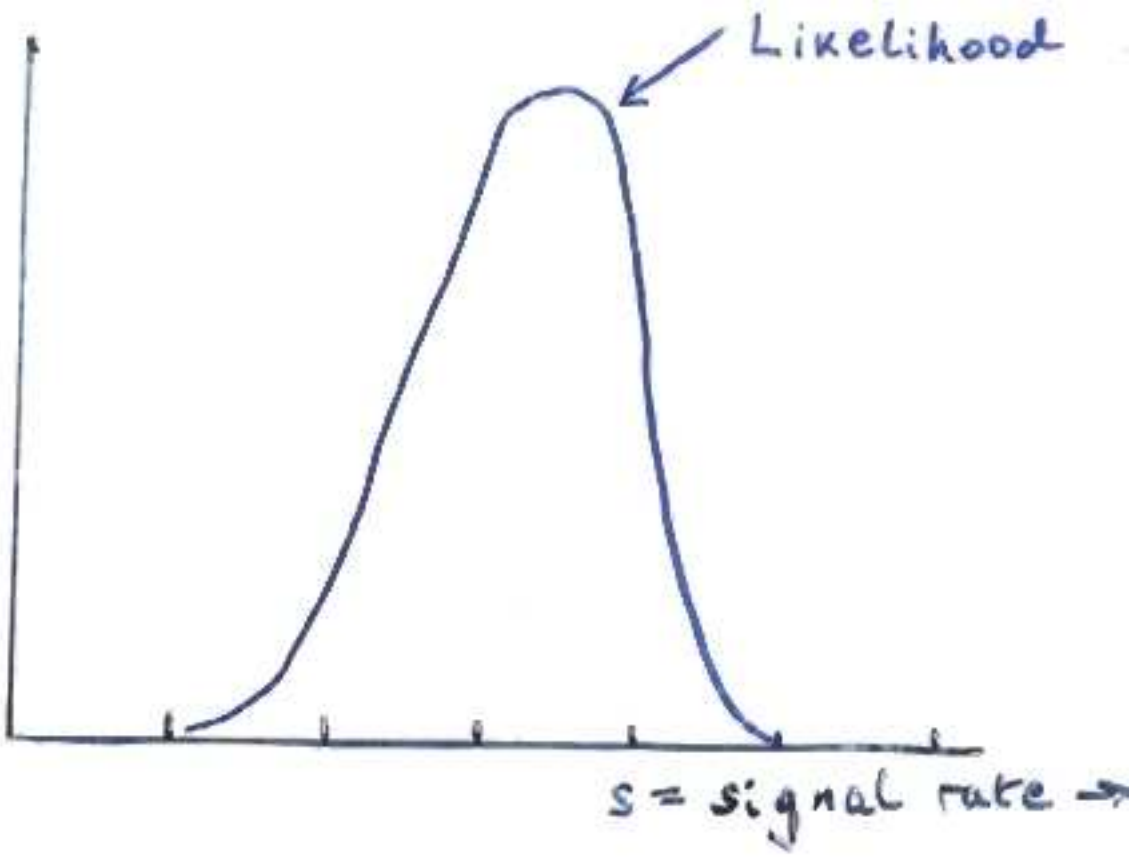
Unimportant if "data overshadows prior"

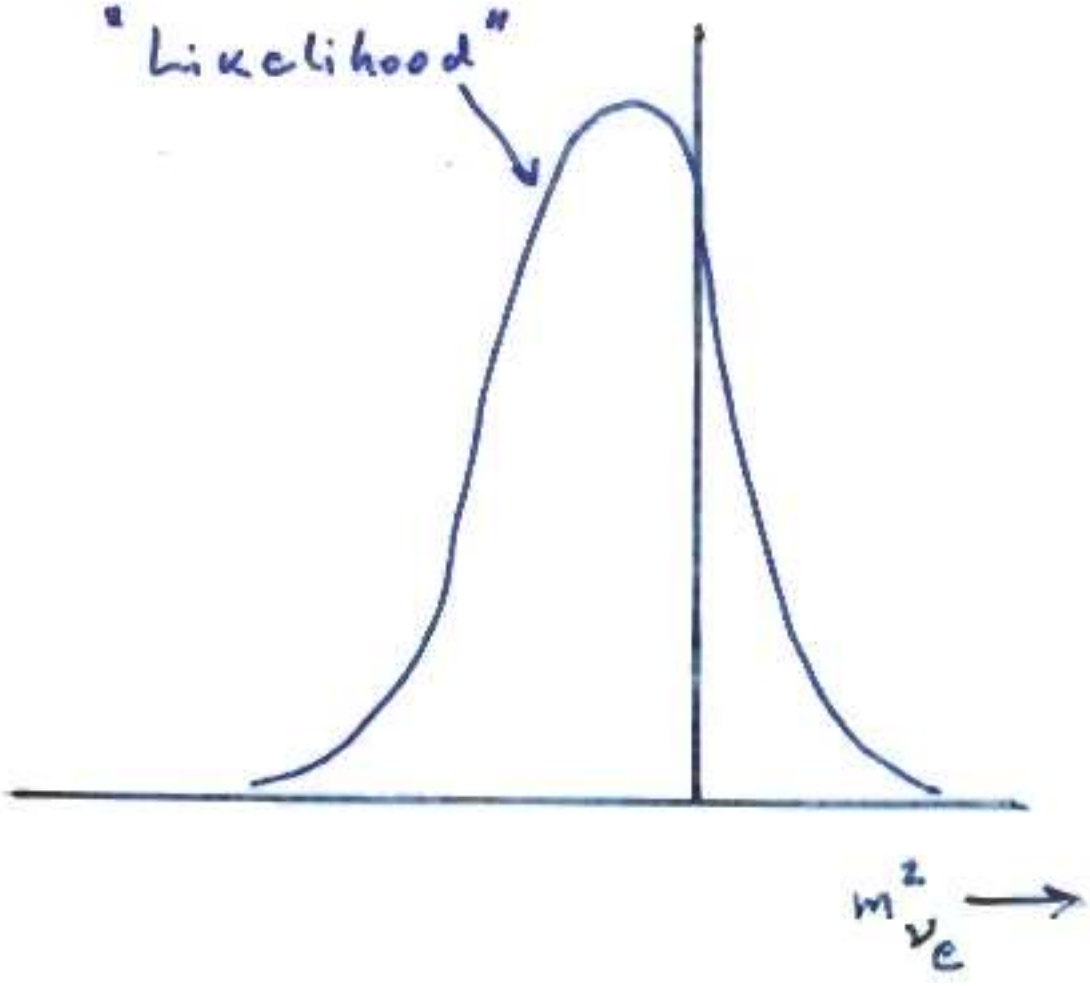
Important for limits

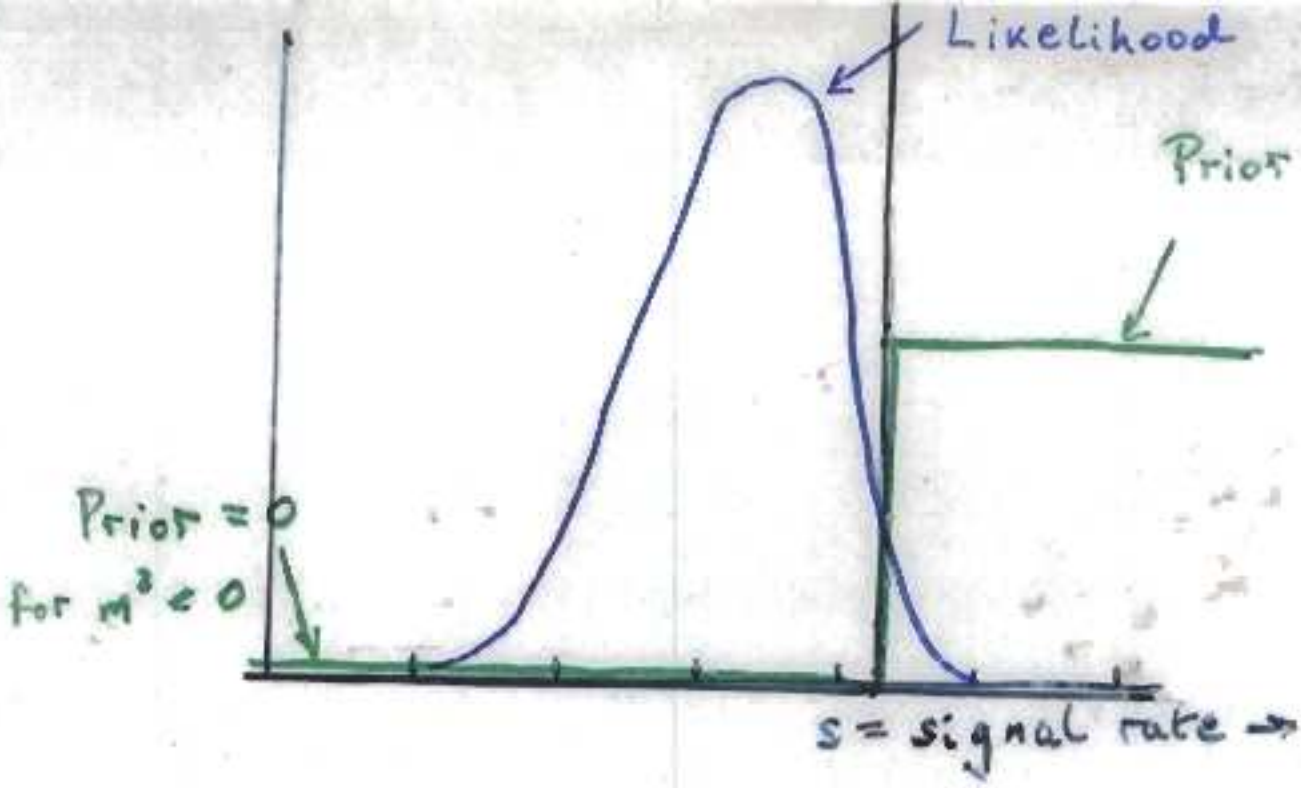
Subjective or Objective prior?

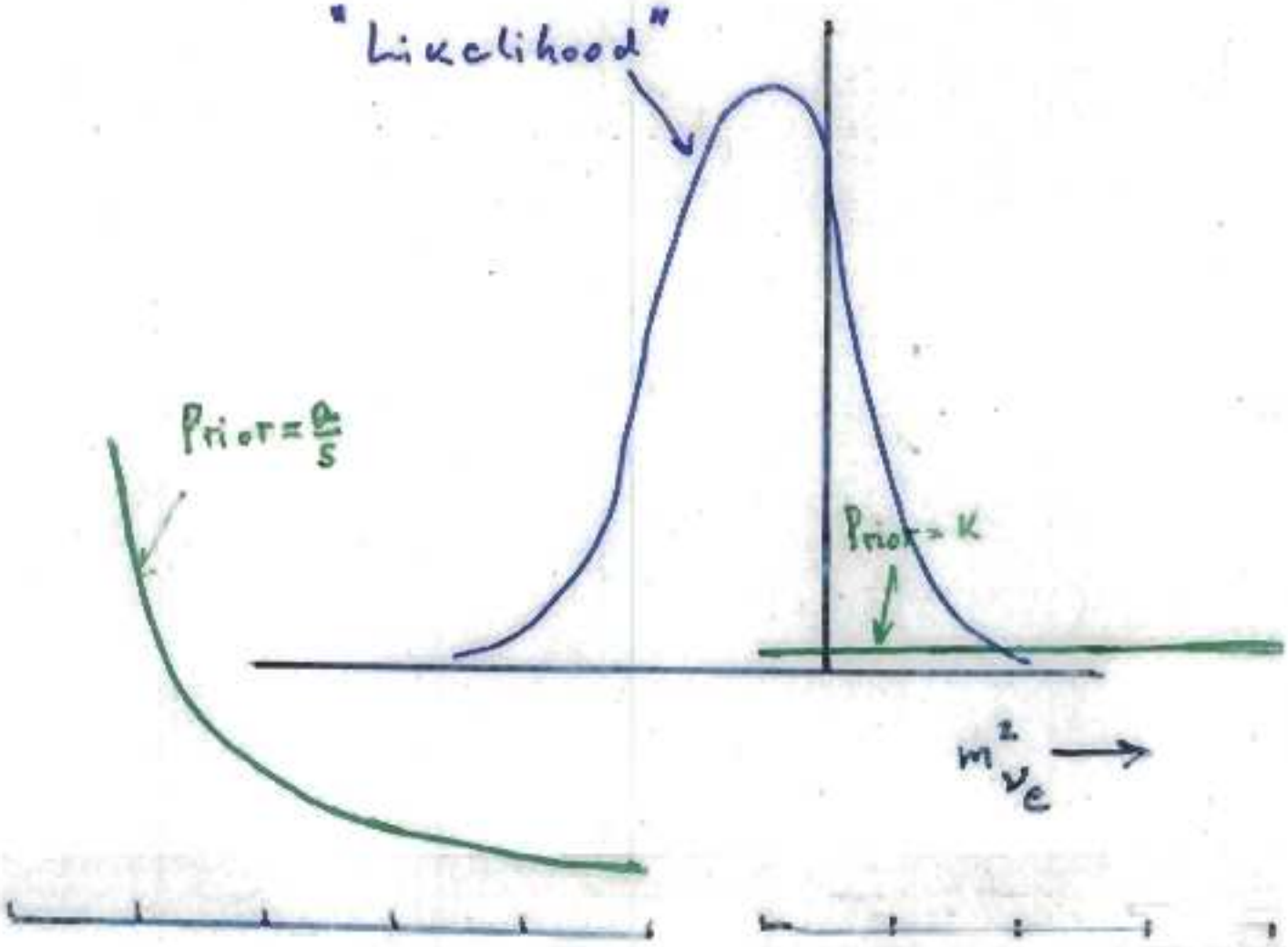


Data overshadows the Prior

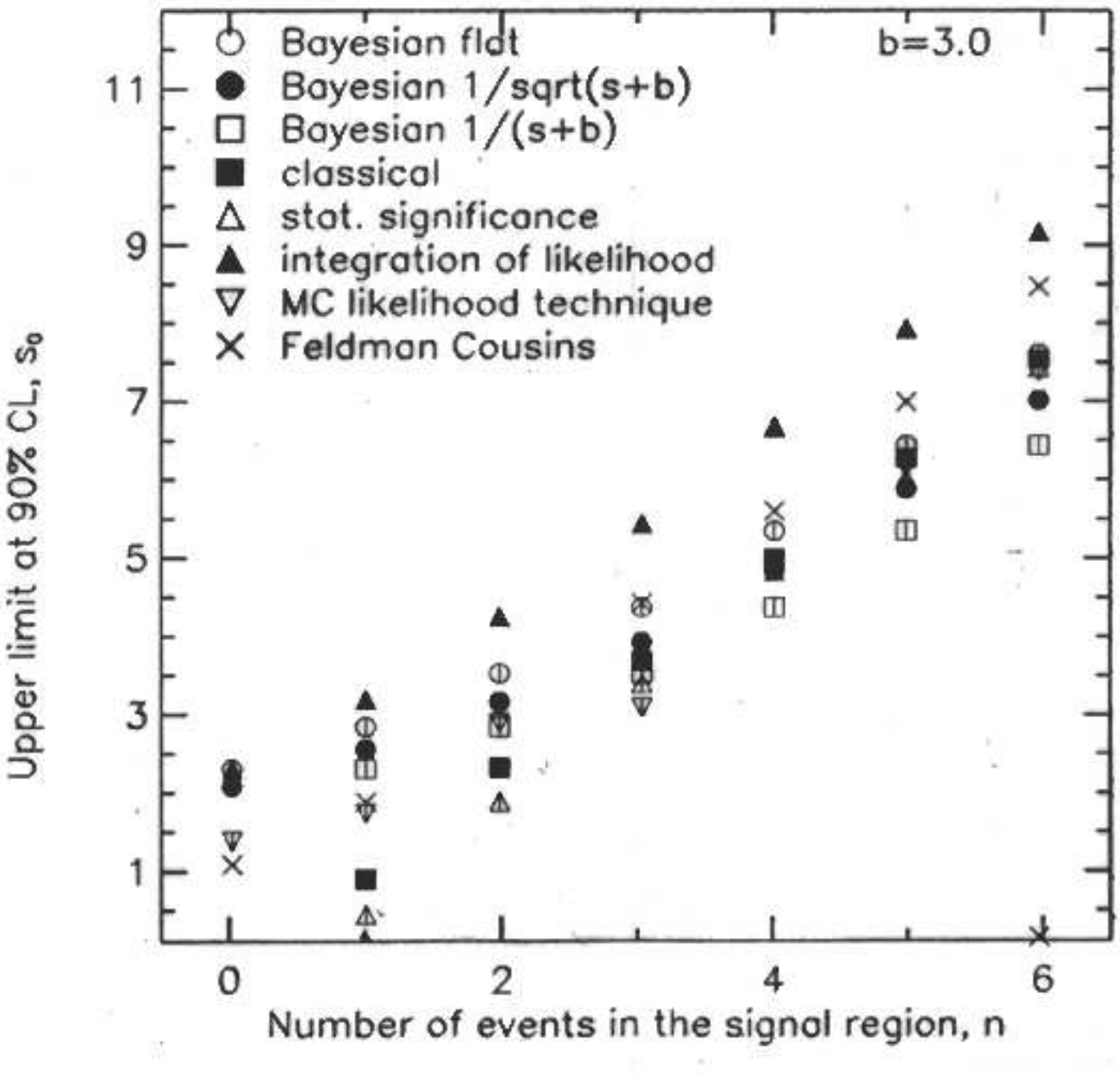

Likelihood
s = signal rate →


"Likelihood"
$m^2_{\nu_e}$ →


Likelihood
Prior
Prior = 0
for $m^2 < 0$
s = signal rate →


"Likelihood"
Prior = $\frac{a}{s}$
Prior = K
$m^2_{\nu_e}$ →

ILYA NARSKY
FNAL CL Workshop



$P(\text{Data};\text{Theory}) \neq P(\text{Theory};\text{Data})$

Theory = male or female

Data = pregnant or not pregnant

$P(\text{pregnant} ; \text{female}) \sim 3\%$

but

$P(\text{female} ; \text{pregnant}) >>> 3\%$

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

HIGGS SEARCH at CERN

Is data consistent with Standard Model?

or with Standard Model + Higgs?

End of Sept 2000 Data not very consistent with S.M.

Prob (Data ; S.M.) < 1% valid frequentist statement

Turned by the press into: Prob (S.M. ; Data) < 1%

and therefore Prob (Higgs ; Data) > 99%

i.e. "It is almost certain that the Higgs has been seen"

Example 1 : Is coin fair ?

Toss coin: 5 consecutive tails

What is P(unbiased; data) ? i.e. p = ½

Depends on Prior(p)

If village priest prior ~ $\delta(1/2)$

If stranger in pub prior ~ 1 for 0<p<1

(also needs cost function)

Example 2 : Particle Identification

Try to separate π and protons

probability (p tag;real p) = 0.95

probability (π tag; real p) = 0.05

probability (p tag ; real (π) = 0.10

probability (π tag ; real π) = 0.90

Particle gives proton tag. What is it?

Depends on prior = fraction of protons

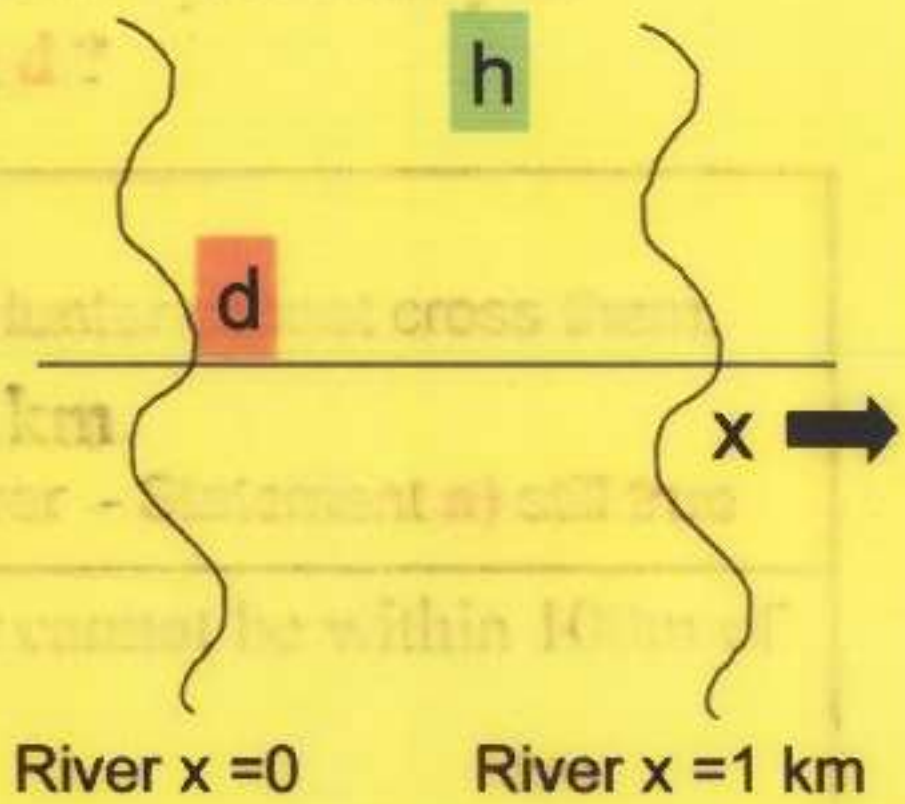
If proton beam, very likely

If general secondary particles, more even

If pure π beam, ~ 0

Hunter and Dog

1) Dog d has 50% probability of being 100 m. of Hunter h
2) Hunter h has 50% probability of being within 100m of Dog d



Given that: a) Dog d has 50% probability of being 100 m. of Hunter

Is it true that b) Hunter h has 50% probability of being within 100m of Dog d ?

Additional information

- Rivers at zero & 1 km. Hunter cannot cross them.

$$0 \le h \le 1\,\text{km}$$

- Dog can swim across river - Statement a) still true

If dog at –101 m, hunter cannot be within 100m of dog

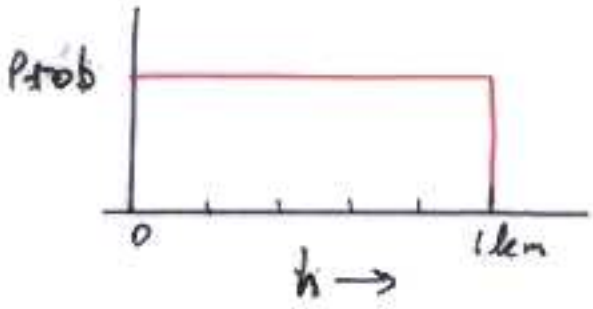
Statement b) untrue

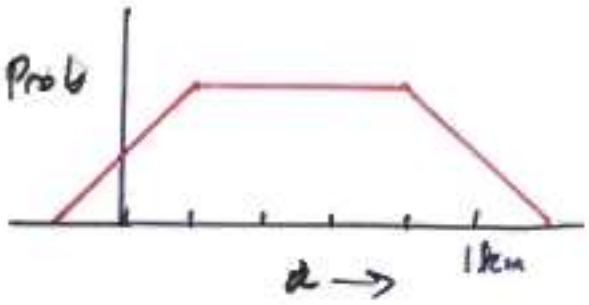
Example:

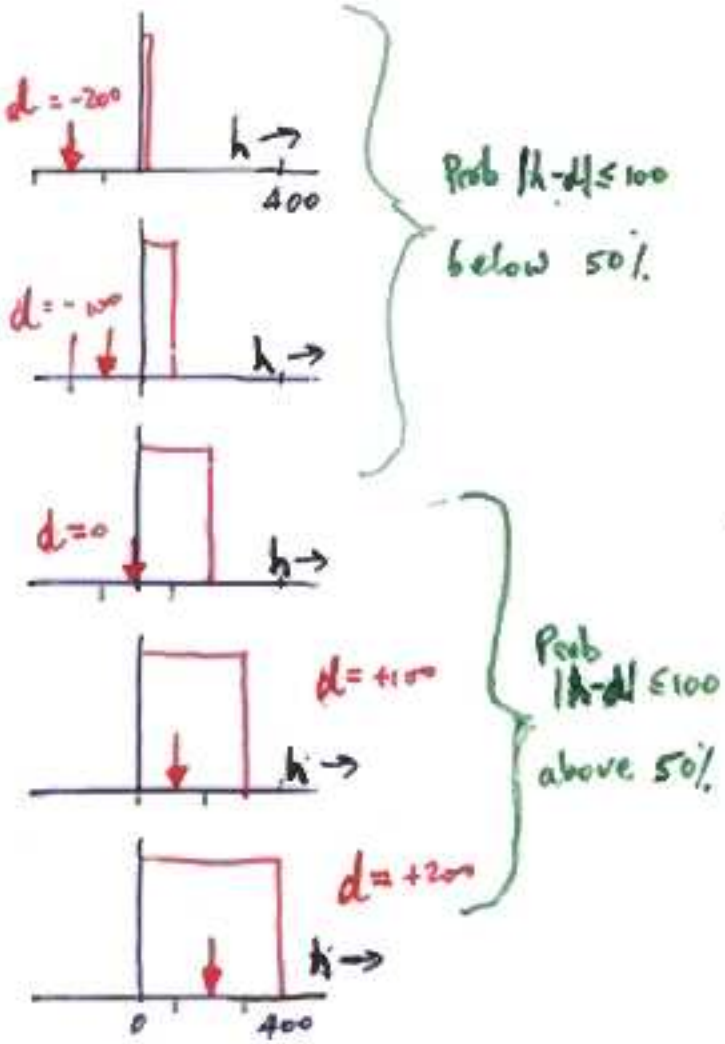
1) More specific on statement ①:

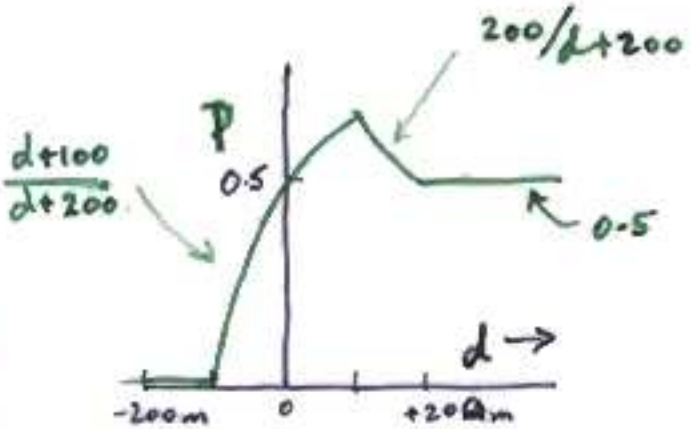
$$\text{Prob}(d-h) = \begin{cases} \text{const} & \text{for } |d-h| < 200\,m \\ 0 & \text{for } |d-h| > 200\,m \end{cases} \quad [\text{L'Hoo}]$$

2) Hunter h uniform in $0 \to 1\,km$ [PRIOR]









P = prob $|h-d| \leq 100\,m$

Classical Approach

Neyman "confidence interval" avoids pdf for μ

uses only P(x; μ)

Confidence interval $\mu_1 \rightarrow \mu_2$:

P($\mu_1 \rightarrow \mu_2$ contains μ) = α True for any μ

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 is in $\mu_1 \rightarrow \mu_2$

CLASSICAL (NEYMAN) CONFIDENCE INTERVALS

Uses only P(data | theory)

FIGURES

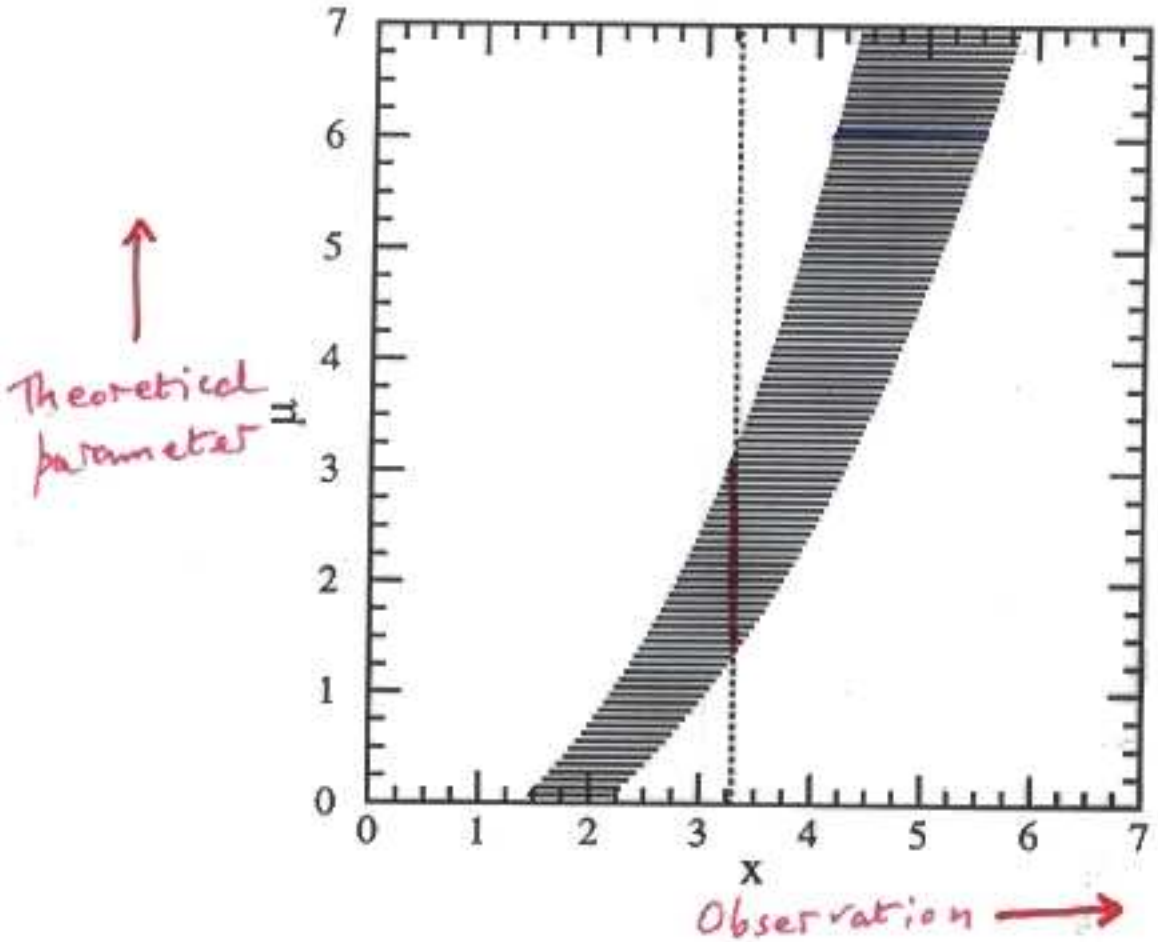


FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2]|\mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0.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

NO PRIOR INVOLVED

BF8

90% classical interval for Gaussian

$\sigma = 1$

$\mu \geq 0$

e.g. $m^2(\nu_e)$

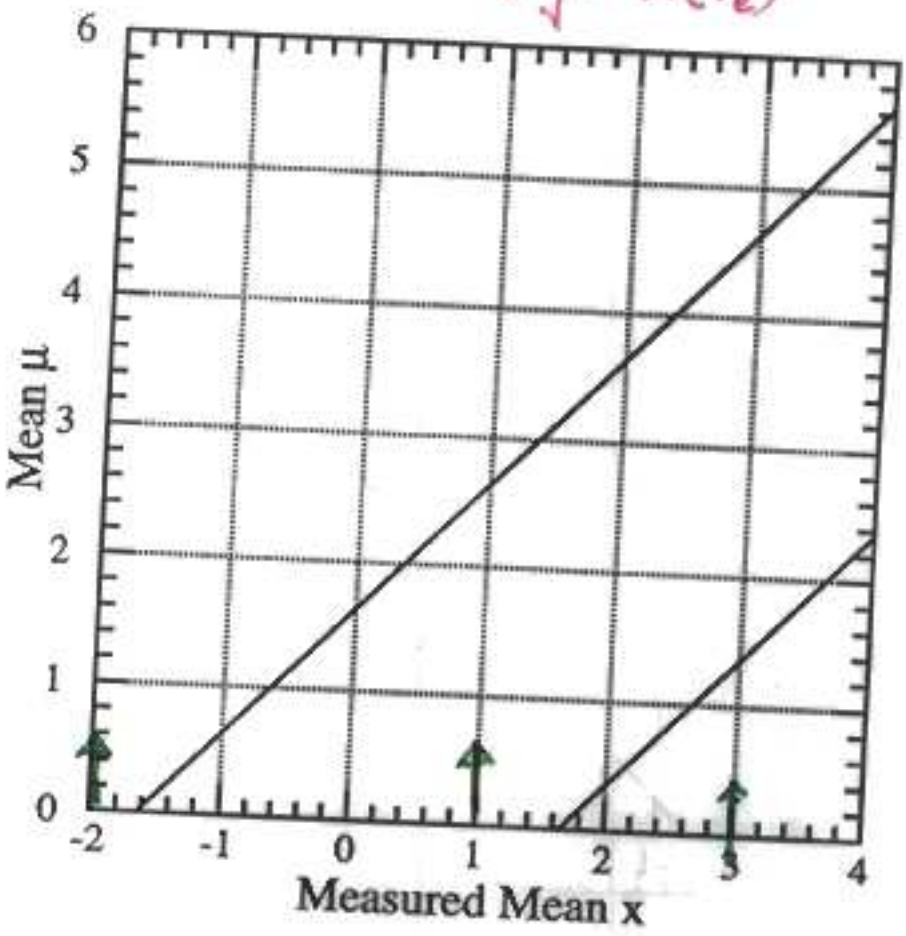


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

$x_{obs} = -2$ No region for μ

RF 10

Classical Intervals

- Problems

Hard to understand e.g. d'Agostini e-mail
Arbitrary choice of interval
Possibility of empty range
Over-coverage for integer observation
e.g. # of events
Nuisance parameters (systematic errors)

- Advantages

Widely applicable
Well defined coverage

$\mu_l \leq \mu \leq \mu_u$ 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

FELDMAN - COUSINS

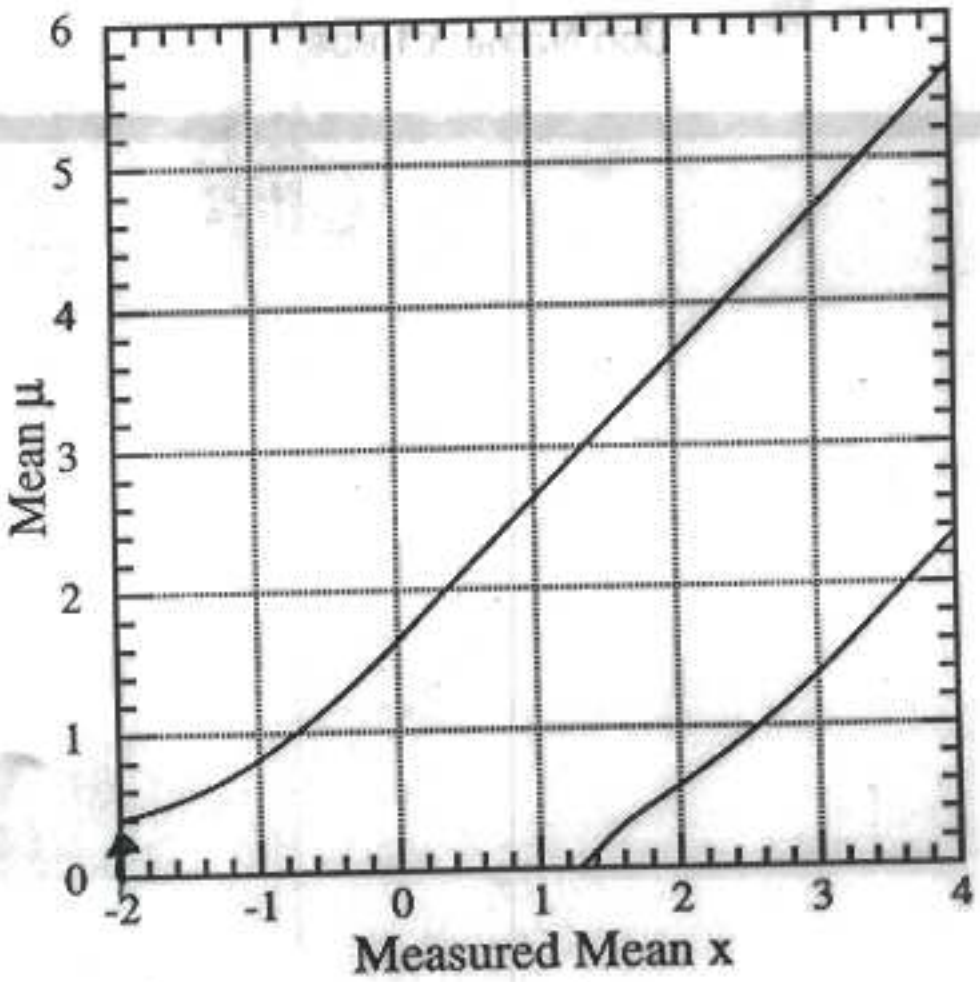
WANT TO AVOID EMPTY CLASSICAL INTERVALS ⇒

USE "$\mathcal{L}$ RATIO ORDERING PRINCIPLE"

TO RESOLVE AMBIGUITY ABOUT "WHICH 90% REGION?" ⇒

[NEYMAN-PEARSON SAY $\mathcal{L}$ RATIO IS BEST FOR HYPOTHESIS TESTING]

NO FLIP-FLOP PROBLEM ⇒



Feldman-Cousins 90% Conf interval

FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

$x_{obs} = -2$ Now gives upper limit

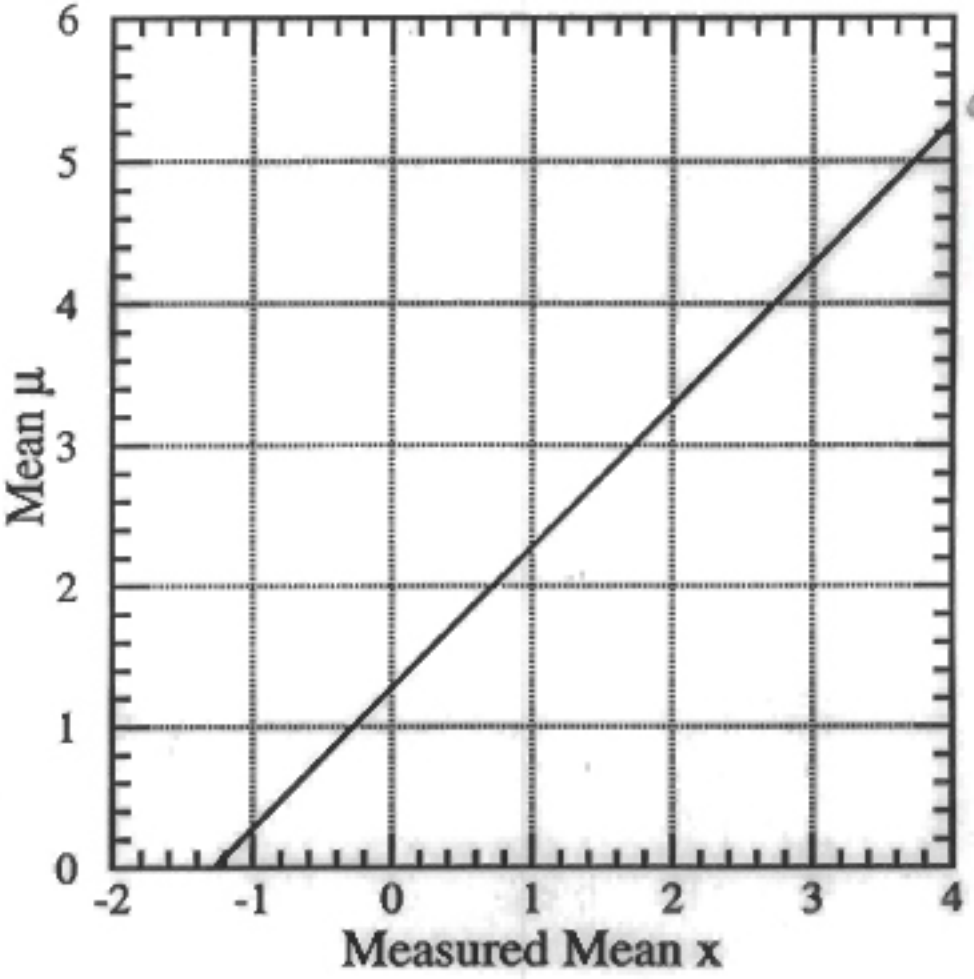


FIG. 2. Standard confidence belt for 90% C.L. upper limits for the mean of a Gaussian, in units of the rms deviation. The second line in the belt is at $x = +\infty$.

FLIP – FLOP

90% upper limit for $x_{obs} \leq 3$

90% 2-sided interval for $x_{obs} > 3$

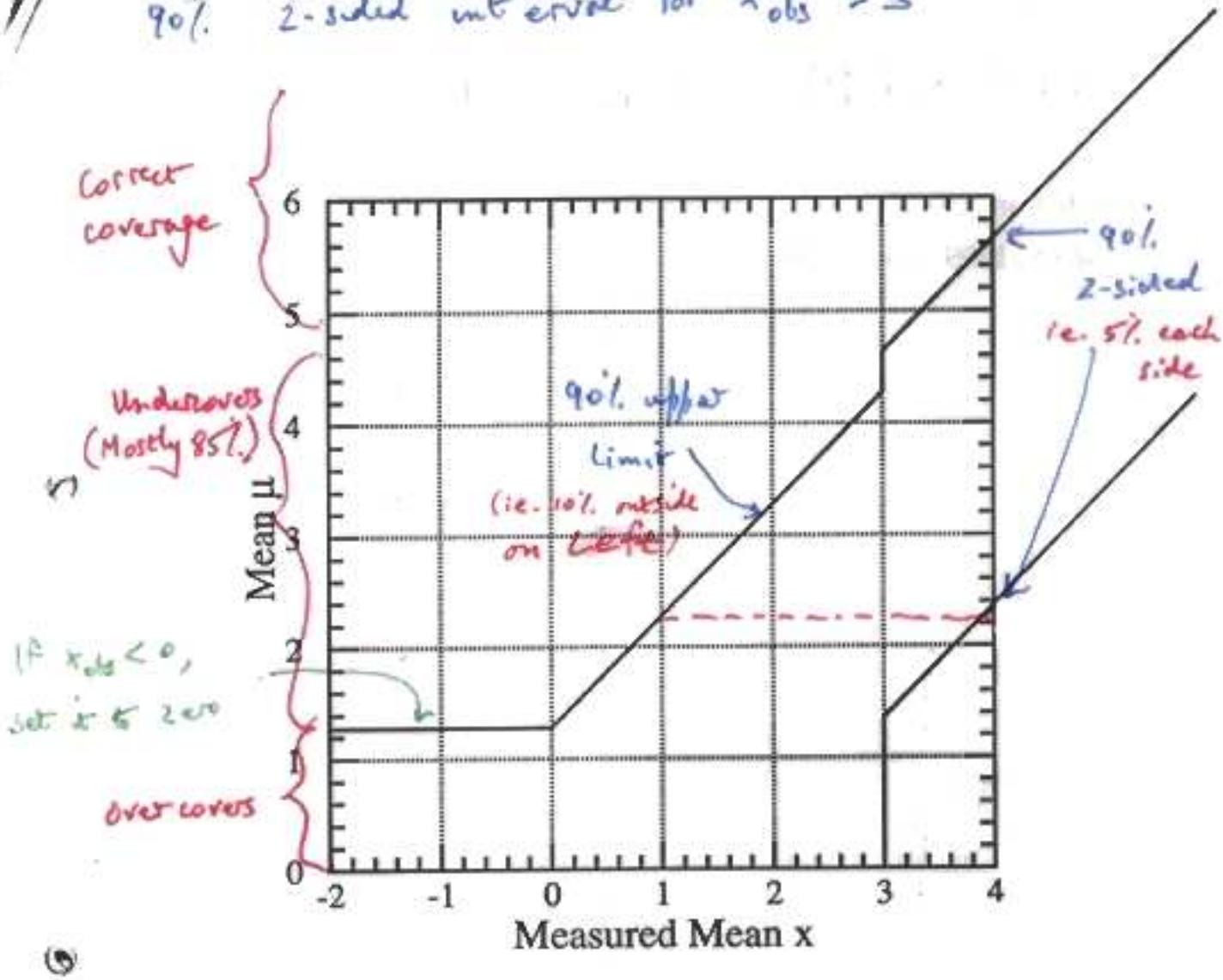


FIG. 4. Plot of confidence belts implicitly used for 90% C.L. confidence intervals (vertical intervals between the belts) quoted by flip-flopping Physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence level. For $1.36 < \mu < 4.28$, the coverage (probability contained in the horizontal acceptance interval) is 85%.

Not good to let x_{obs} determine how result will be presented

F-C goes smoothly from 1-sided → 2-sided

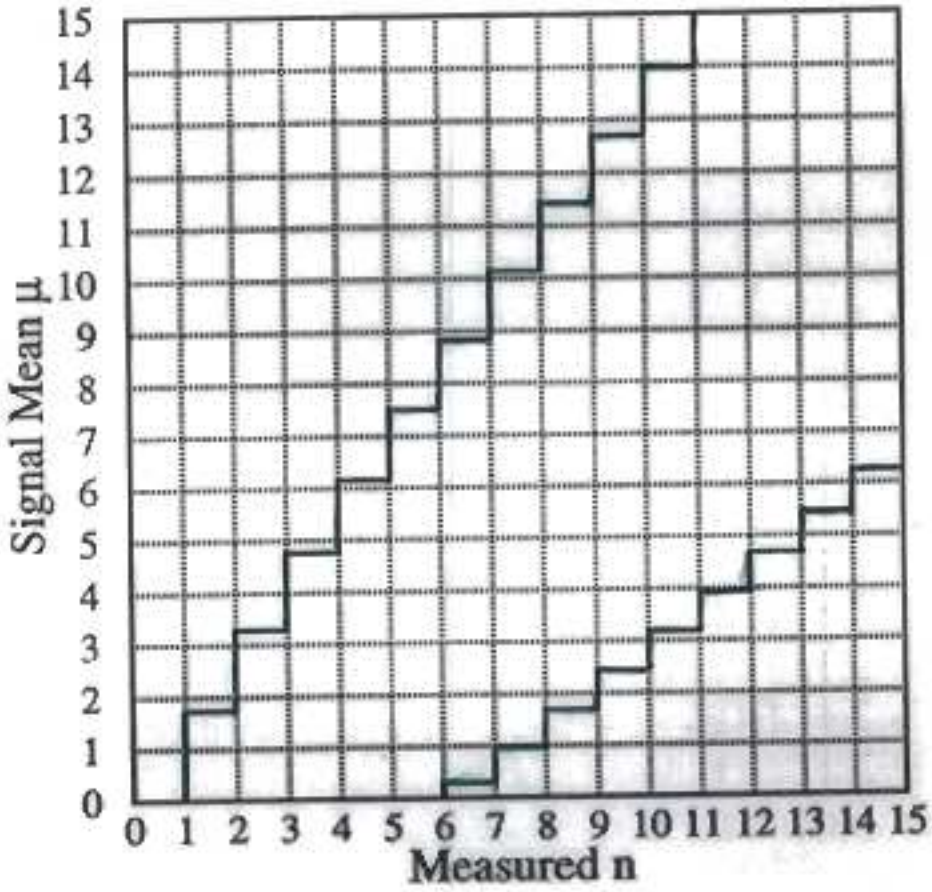


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist
for Poisson mean μ

FELDMAN + COUSINS FOR

POISSON MEAN μ

90% Conf

$b = 3.0$

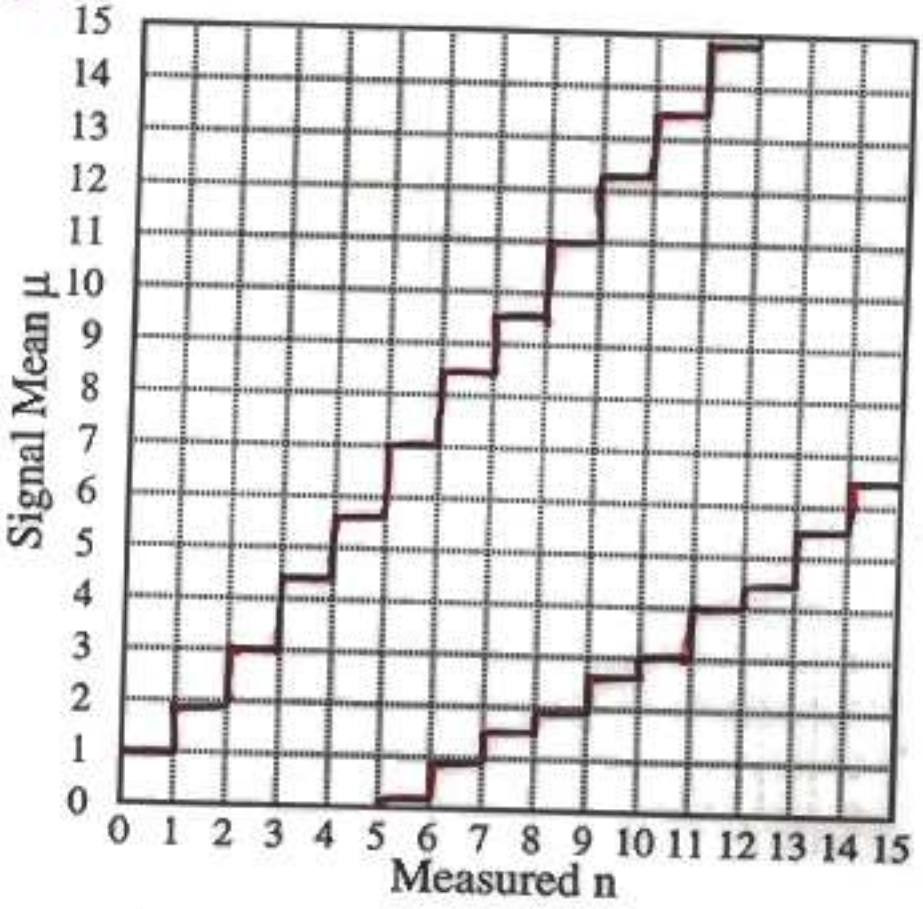


FIG. 7. Confidence belt based on our ordering principle, for 90% C.L. confidence intervals for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

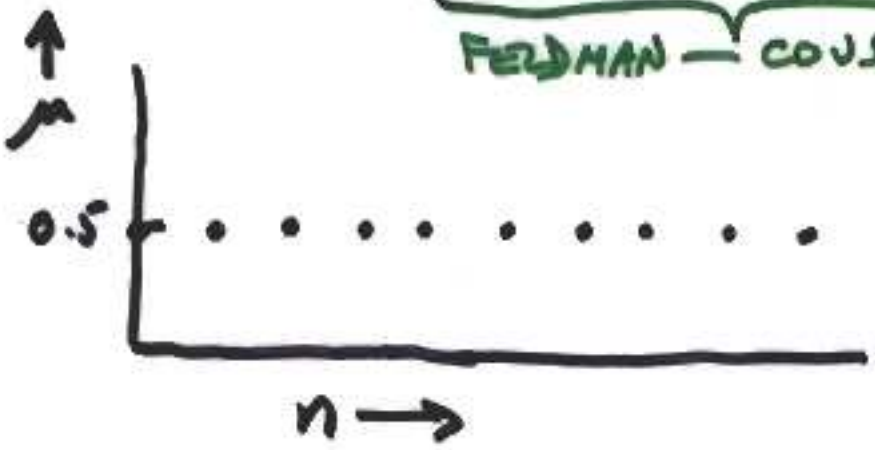
FREQUENTIST POISSON C.B. CONSTRN.

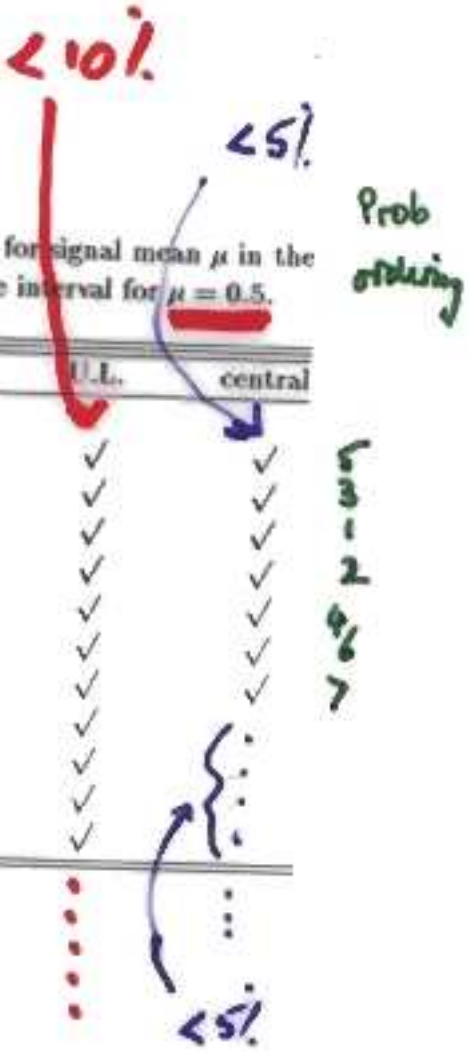
TABLES

TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

n	$P(n\|\mu)$	μ_{best}	$P(n\|\mu_{\text{best}})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	✓	✓
2	0.185	0.	0.224	0.826	3	✓	✓
3	0.216	0.	0.224	0.963	2	✓	✓
4	0.189	1.	0.195	0.966	1	✓	✓
5	0.132	2.	0.175	0.753	4	✓	✓
6	0.077	3.	0.161	0.480	7	✓	✓
7	0.039	4.	0.149	0.259		✓	✓
8	0.017	5.	0.140	0.121		✓	
9	0.007	6.	0.132	0.050		✓	
10	0.002	7.	0.125	0.018		✓	
11	0.001	8.	0.119	0.006		✓	

FELDMAN — COUSINS



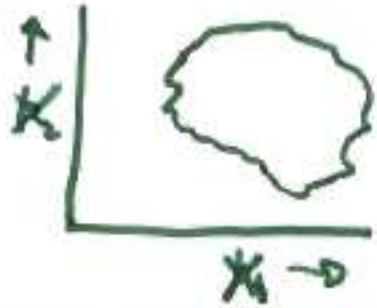


FEATURES OF F+C

REDUCES EMPTY INTERVALS

- UNIFIED 1-SIDED & 2-SIDED INTERVALS
- ELIMINATES FLIP-FLOP
- NO ARBITRARINESS OF INTERVAL

"READILY" EXTENDS TO SEVERAL DIMENSIONS



LESS OVERCOVERAGE THAN "5% AT ENDS"

MAY PROB DENSITY 5% AT ENDS ?

NEYMAN CONSTRUCTION ⇒ CPU-INTENSIVE (ESP IN SEVERAL DIMENSIONS)

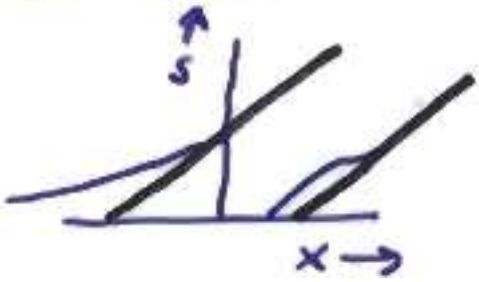
MINOR PATHOLOGIES : DISTANT INTERVALS

WRONG BEHAVIOUR WRT BG

TIGHT LIMITS FOR $b > n_{obs}$ e.g.

n_{obs}	b_{gal}	90% Limit
0	3.0	1.08
0	0	2.44

UNIFIED ⇒ QUICKER EXCLUSION OF $s=0$



Bayesian

Pros:

Easy to understand

Physical Interval

Cons:

Needs prior

Hard to combine

Coverage

Standard Frequentist

Pros:

Coverage

Cons:

Hard to understand

Small or Empty Intervals

Different Upper Limits

Bayesian versus Frequentism

	Bayesian	Frequentist
Basis of method	Bayes Theorem --> Posterior probability distribution	Uses pdf for data, for fixed parameters
Meaning of probability	Degree of belief	Frequentist defintion
Prob of parameters?	Yes	Anathema
Needs prior?	Yes	No
Choice of interval?	Yes	Yes (except F+C)
Data considered	Only data you have	….+ more extreme
Likelihood principle?	Yes	No

Bayesian versus Frequentism

	Bayesian	Frequentist
Ensemble of experiments	No	Yes (but often not explicit)
Final statement	Posterior probability distribution	Parameter values → Data is likely
Unphysical/ empty ranges	Excluded by prior	Can occur
Systematics	Integrate over prior	Extend dimensionality of frequentist construction
Coverage	Unimportant	Built-in
Decision making	Yes (uses cost function)	Not useful

Bayesianism versus Frequentism

"Bayesians address the question everyone is interested in, by using assumptions no-one believes"

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

PRACTICAL STATISTICS #5

Hunter + Dog paradox

Bayes + Frequentism, contd
2 more Bayes examples
Frequentist approach

} See #4

Feldman – Cousins
Ordering rule
Flip – flop
Gaussian + Poisson examples
ν oscillations

Systematics
Shift method
Profile $\mathcal{L}$
Bayes
Frequentist
Mixed: Cousins + Highland

Multivariate analysis
Neural networks

BLUE combination technique

LOUIS LYONS
CDF

FELDMAN - COUSINS

WANT TO AVOID EMPTY CLASSICAL INTERVALS ⇒

USE "$\mathcal{L}$ RATIO ORDERING PRINCIPLE"

TO RESOLVE AMBIGUITY ABOUT "WHICH 90% REGION?" ⇒

[NEYMAN-PEARSON SAY $\mathcal{L}$ RATIO IS BEST FOR HYPOTHESIS TESTING]

NO FLIP-FLOP PROBLEM ⇒

90% classical interval for Gaussian

$\sigma = 1$

$\mu \geq 0$

e.g. $m^2(\nu_e)$

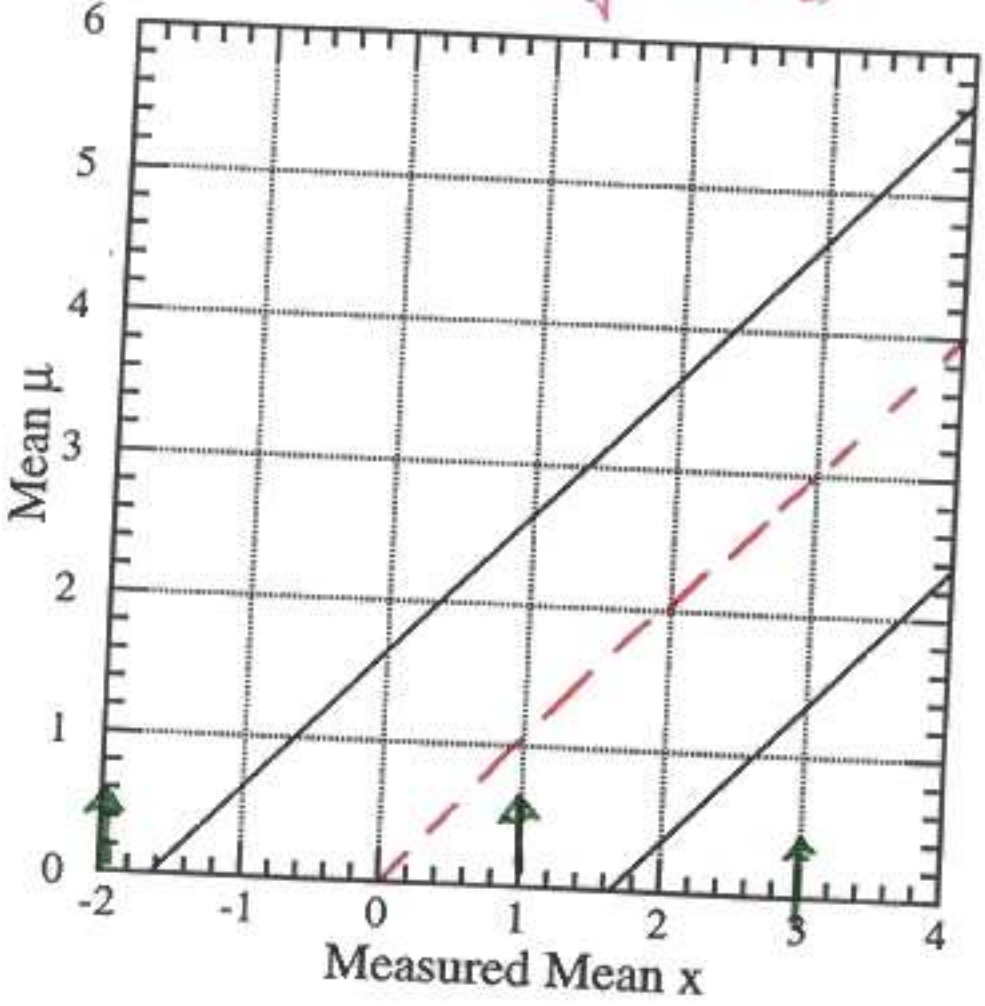


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

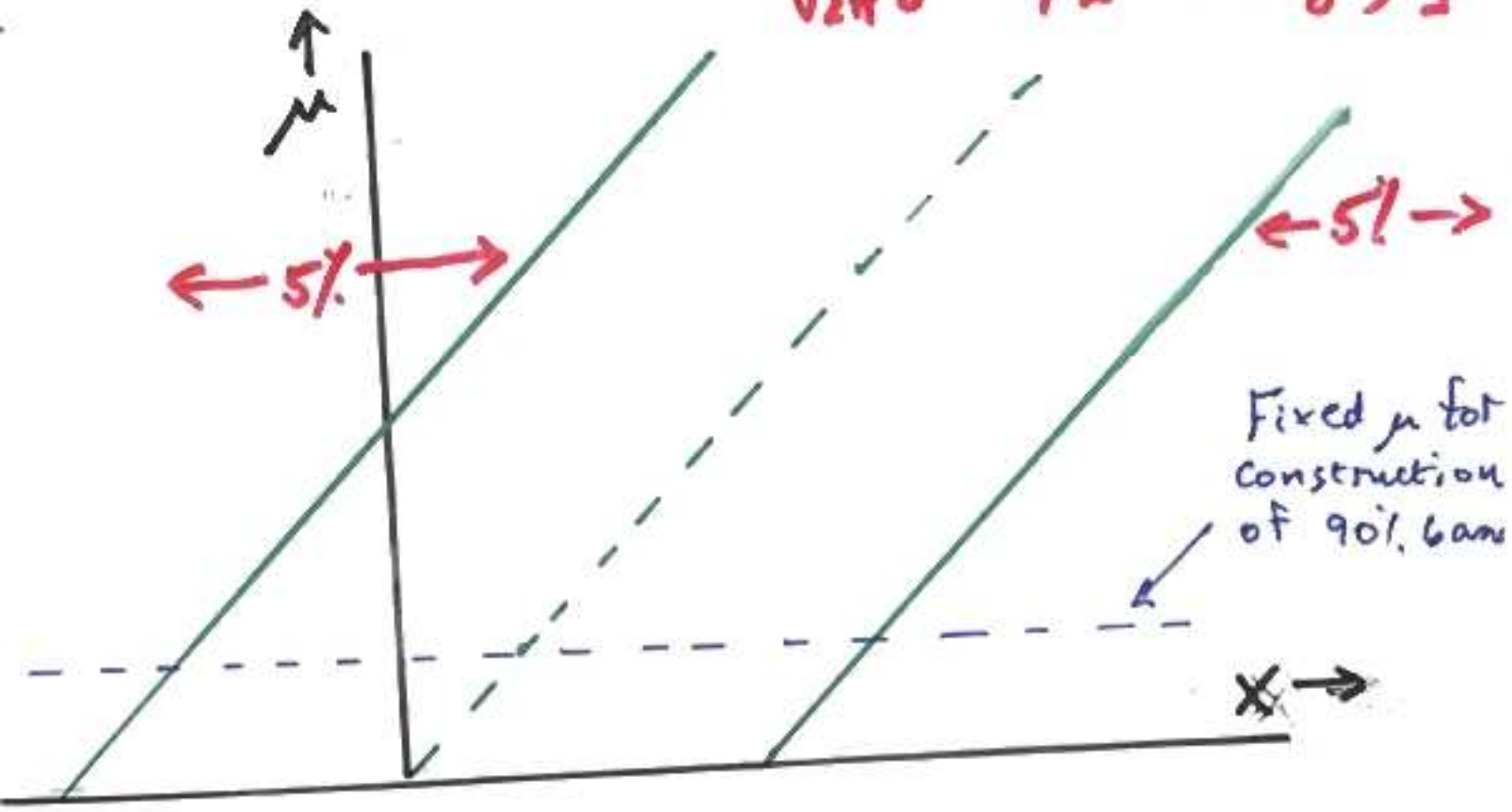
$x_{obs} = -2$ No region for μ

FELDMAN-COUSINS ORDERING RULE

$R = p(x,\mu) / p(x,\mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

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

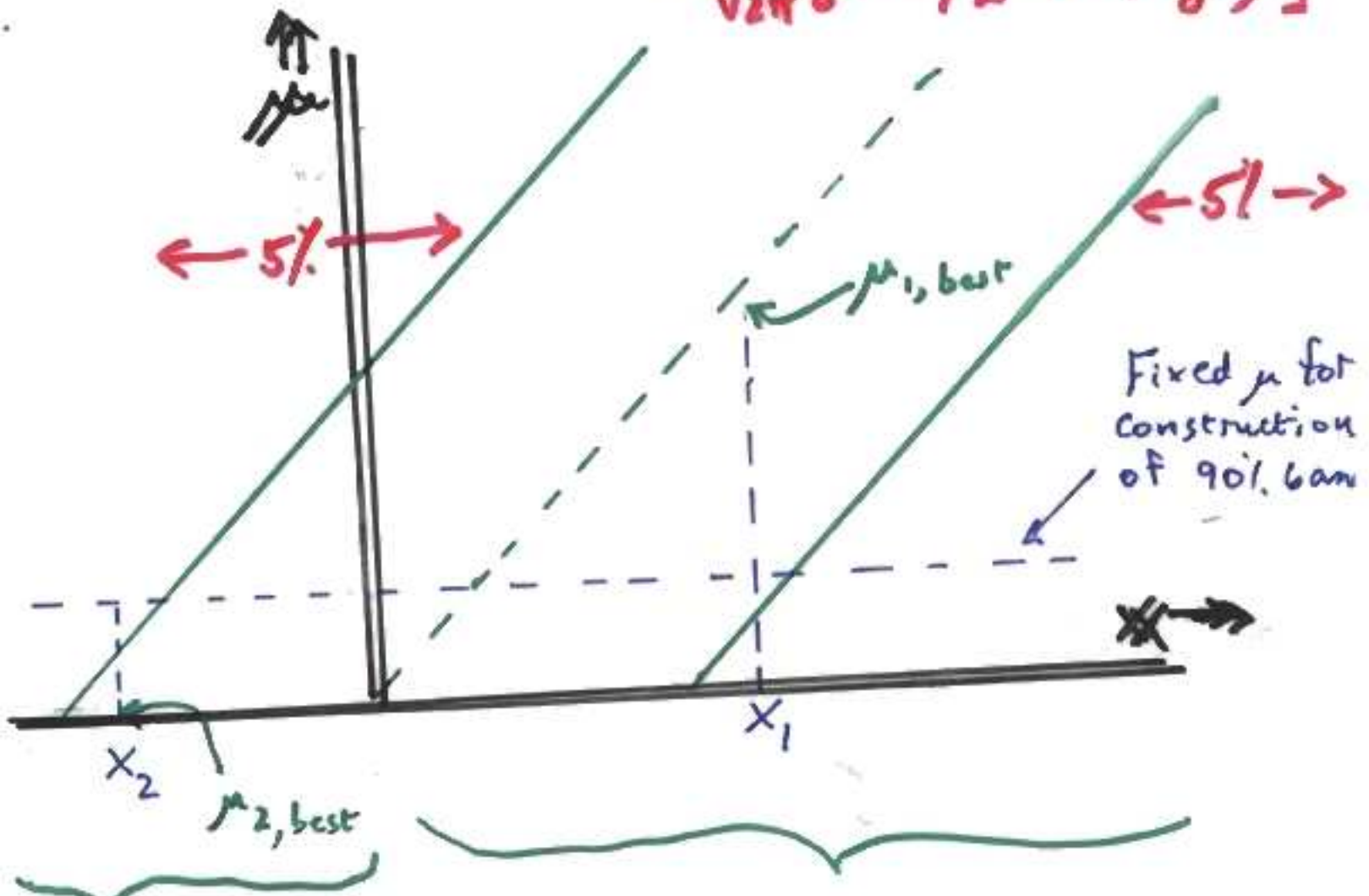


FELDMAN-COUSINS ORDERING RULE

$R = p(x,\mu) / p(x,\mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

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



$\mu_{best} = 0$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2}$

$p(x_1,\mu) > p(x_2,\mu)$

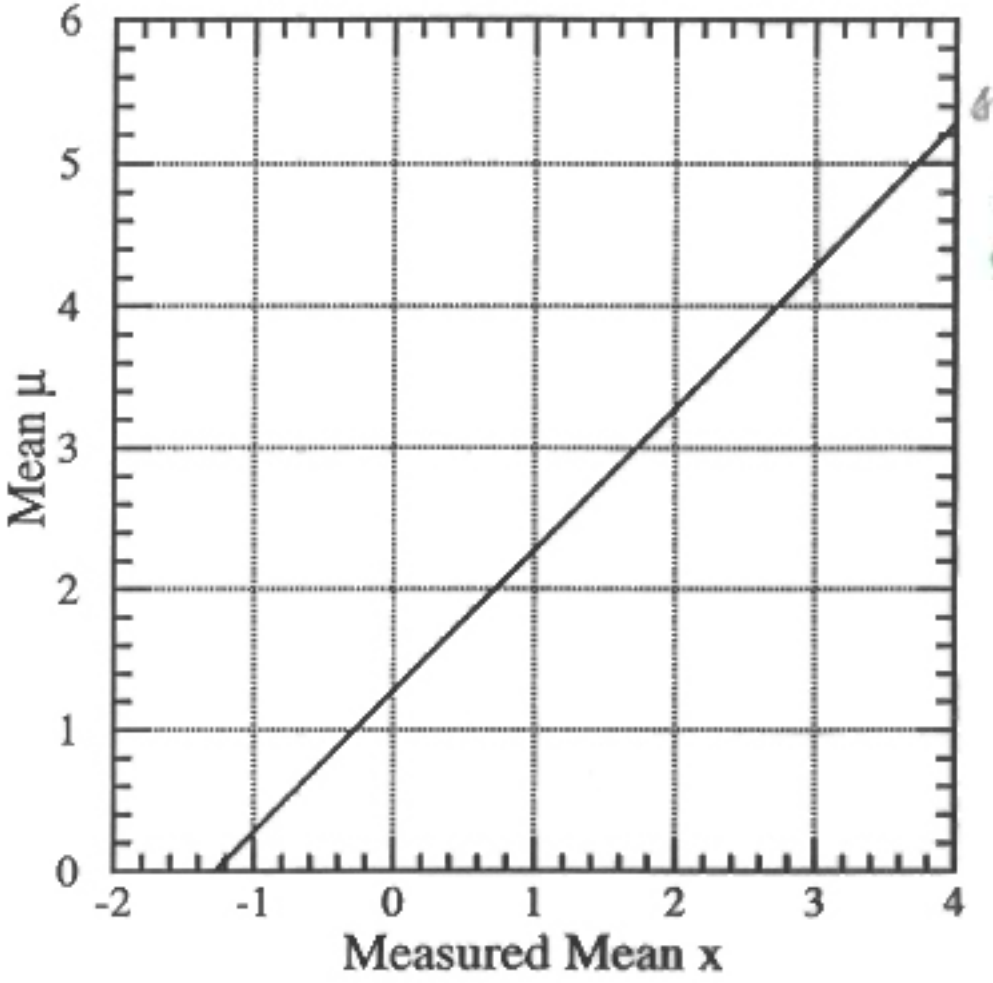
BUT $R(x_2,\mu) > R(x_1,\mu)$

$\mu_{best} = x$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\sigma} = \text{const}$

Standard : Select x_1 before x_2

F-C : Select x_2 before x



classical 90% upper limit

FIG. 2. Standard confidence belt for 90% C.L. upper limits for the mean of a Gaussian, in units of the rms deviation. The second line in the belt is at $x = +\infty$.

Feldman-Cousins 90% Conf interval

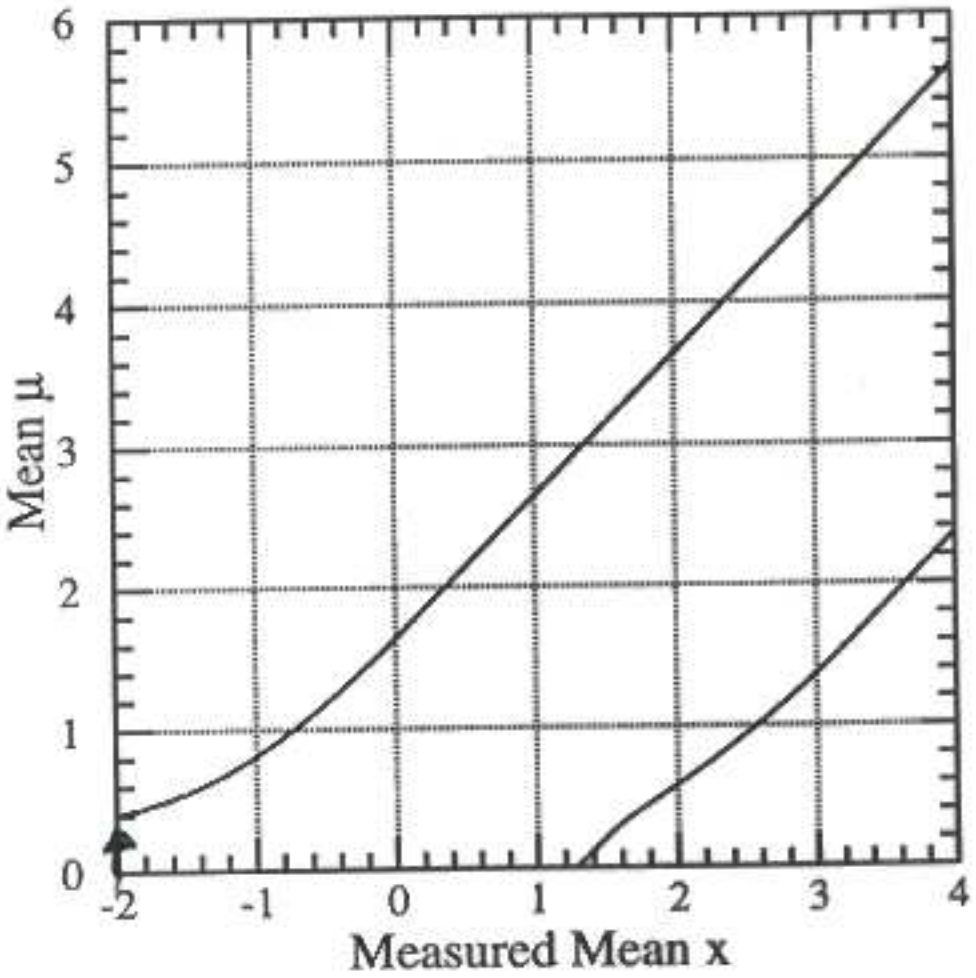


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

$x_{obs} = -2$ Now gives upper limit

FLIP – FLOP

90% upper limit for $x_{obs} \leq 3$

90% 2-sided interval for $x_{obs} > 3$

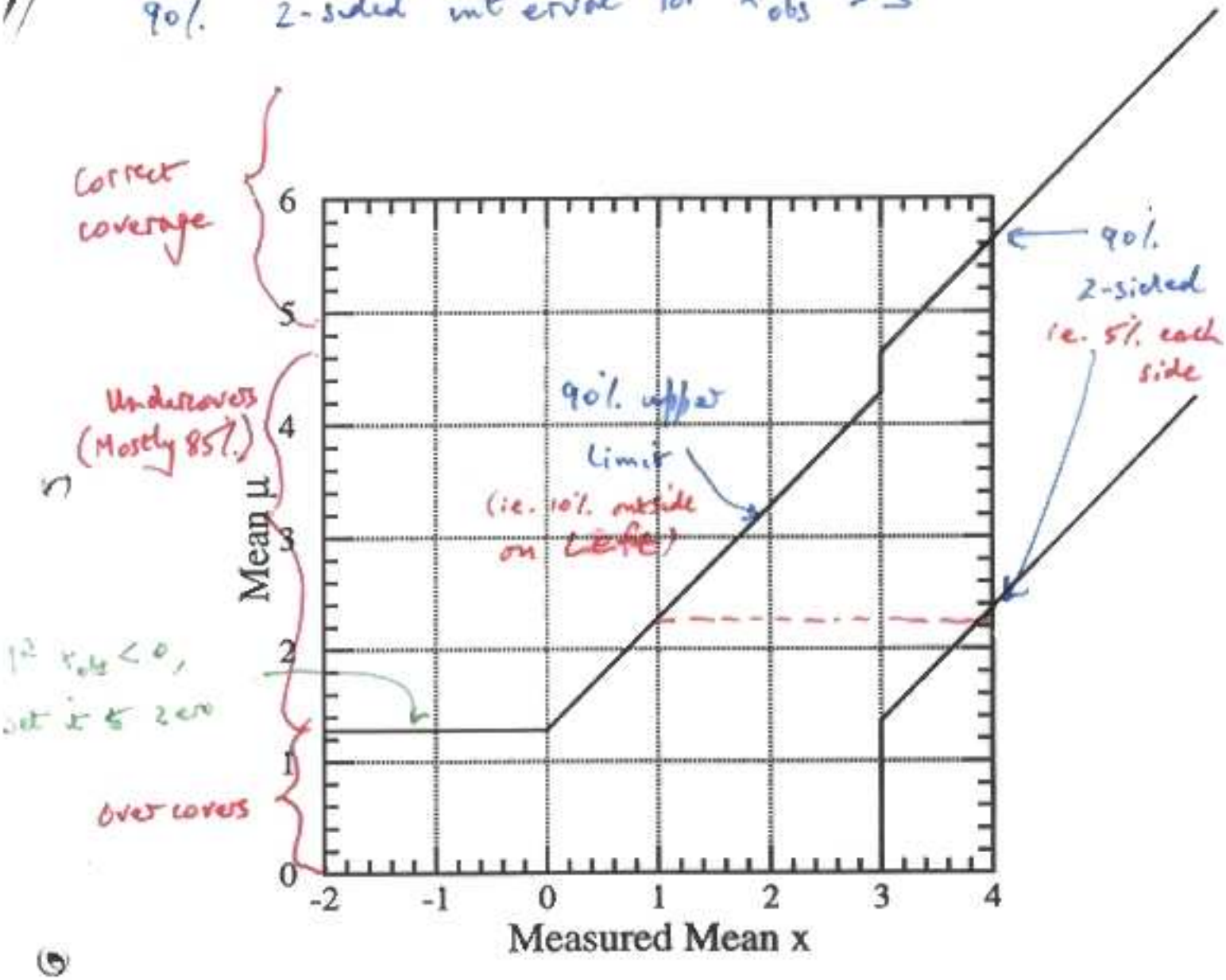


FIG. 4. Plot of confidence belts implicitly used for 90% C.L. confidence intervals (vertical intervals between the belts) quoted by flip-flopping Physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence level. For $1.36 < \mu < 4.28$, the coverage (probability contained in the horizontal acceptance interval) is 85%.

Not good to let x_{obs} determine how result will be presented

F-C goes smoothly from 1-sided → 2-sided

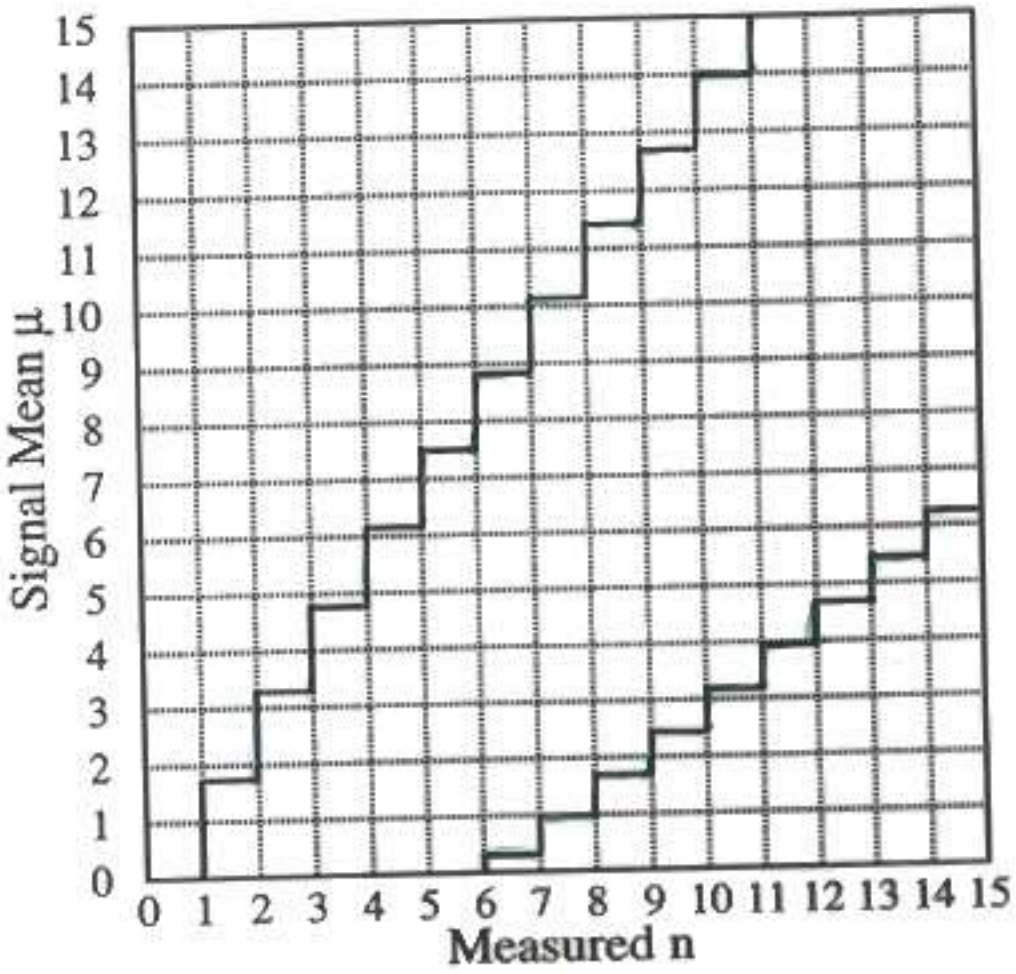


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist

for Poisson mean μ

FELDMAN & COUSINS FOR POISSON MEAN μ

90% Conf

$b = 3.0$

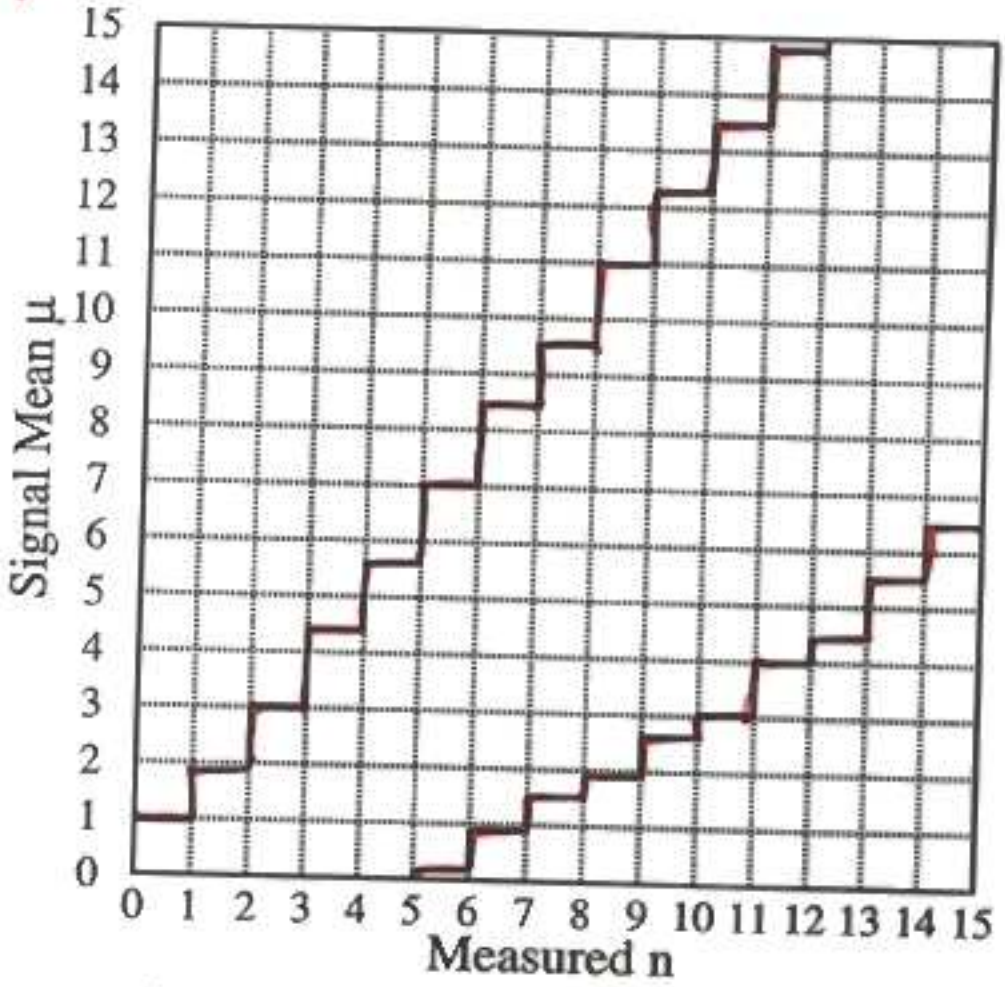


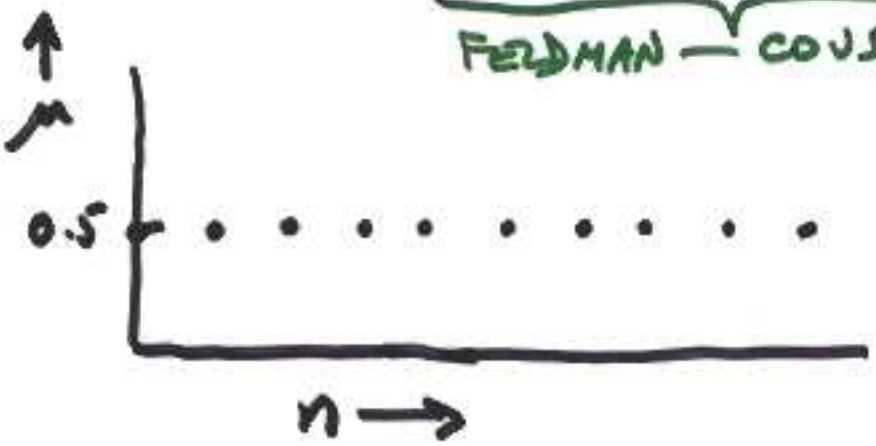
FIG. 7. Confidence belt based on our ordering principle, for 90% C.L. confidence intervals for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

FREQUENTIST POISSON C.B. CONSTR.

TABLES

TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

n	$P(n\|\mu)$	$\mu_{\rm best}$	$P(n\|\mu_{\rm best})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	✓	✓
2	0.185	0.	0.224	0.826	3	✓	✓
3	0.216	0.	0.224	0.963	2	✓	✓
4	0.189	1.	0.195	0.966	1	✓	✓
5	0.132	2.	0.175	0.753	4	✓	✓
6	0.077	3.	0.161	0.480	7	✓	✓
7	0.039	4.	0.149	0.259		✓	✓
8	0.017	5.	0.140	0.121		✓	
9	0.007	6.	0.132	0.050		✓	
10	0.002	7.	0.125	0.018		✓	
11	0.001	8	0.119	0.006		✓	



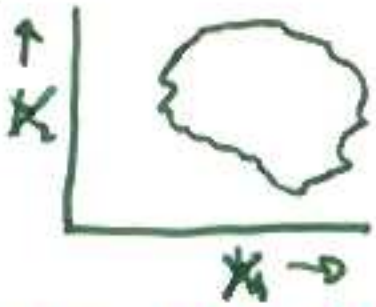
FELDMAN — COUSINS

FEATURES OF F+C

REDUCES EMPTY INTERVALS

- UNIFIED 1-SIDED & 2-SIDED INTERVALS
- ELIMINATES FLIP-FLOP
- NO ARBITRARINESS OF INTERVAL

"READILY" EXTENDS TO SEVERAL DIMENSIONS



LESS OVERCOVERAGE THAN "5% AT ENDS"

MAY PROB DENSITY?
5% AT ENDS ?

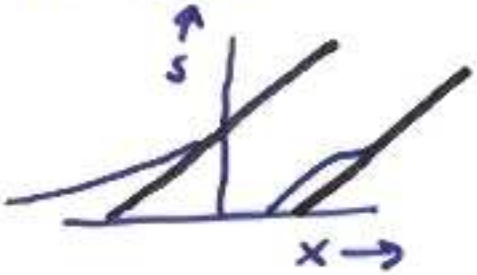
NEYMAN CONSTRUCTION ⇒ CPU-INTENSIVE
(ESP IN SEVERAL DIMENSIONS)

MINOR PATHOLOGIES : DISTANT INTERVALS
WRONG BEHAVIOUR WRT BGD

TIGHT LIMITS FOR $b > n_{obs}$ e.g.

n_{obs}	b_{gal}	90% Limit
0	3.0	1.08
0	0	2.44

UNIFIED ⇒ QUICKER EXCLUSION OF $s=0$

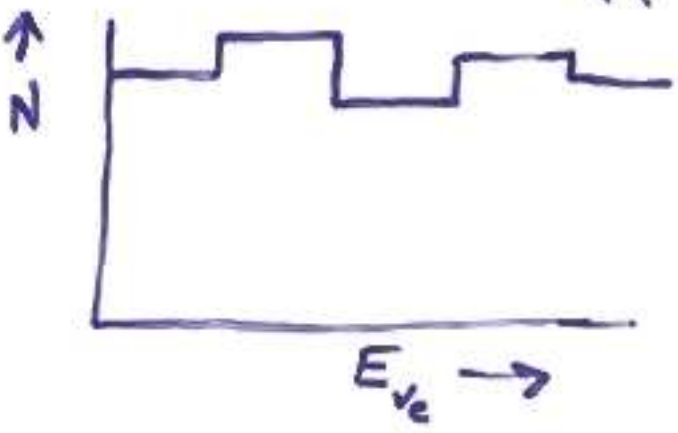


NEUTRINO OSCILLATIONS

$$P(\nu_\mu \to \nu_e) = \sin^2 2\theta \sin^2\left[\frac{1.27\,\Delta m^2 L}{E}\right]$$

Δm^2: eV²; E: GeV; L: km

Data = "ν_e" energy spectrum



Bgd = 100 ev/bin

Signal = 10,000 ev/bin if P = 1

Compare data & prediction via

$$\Delta\chi^2 = \sum \left(\frac{n_i - b_i - \mu_i}{\sigma_i}\right)^2 - \left(\frac{n_i - b_i - (\mu_{best})_i}{\sigma_i}\right)^2$$

OR $$2\sum\left\{\mu_i - (\mu_{best})_i + n_i \frac{\mu_{best,i} + b_i}{\mu_i + b_i}\right\}$$

(Ln [Likelihood ratio])

N.B. $\Delta\chi^2$ is more than just one piece of data

FIND ACCEPTANCE REGION FOR "DATA" BY M.C.

i.e. HOW BIG SHOULD $\Delta\chi^2_{cut}$ BE FOR 90% ACCEPTANCE?

[NOT STANDARD 4.61 for χ^2_2

BECAUSE a) EFFECT OF BOUNDARIES

b) WRONG OVERALL MINIMUM

c) POISSON $\not\equiv$ GAUSSIAN

d) $\Delta\chi^2 \not\equiv \chi^2_2$

e) "1-2" REGIONS AT LOW Δm^2

$[P \sim \sin^2 2\theta \cdot (\Delta m^2)^2]$

$\Delta\chi^2_{cut} = 2.4 - 6.6$

FINALLY, USE DATA $\Rightarrow \Delta\chi^2$ AT EACH $(\sin^2 2\theta, \Delta m^2)$ & COMPARE WITH $\Delta\chi^2_{cut}(\sin^2 2\theta, \Delta m^2)$

TO FIND ACCEPTABLE REGION IN $(\sin^2 2\theta, \Delta m^2)$

VERY MUCH BETTER THAN "RASTER SCAN"

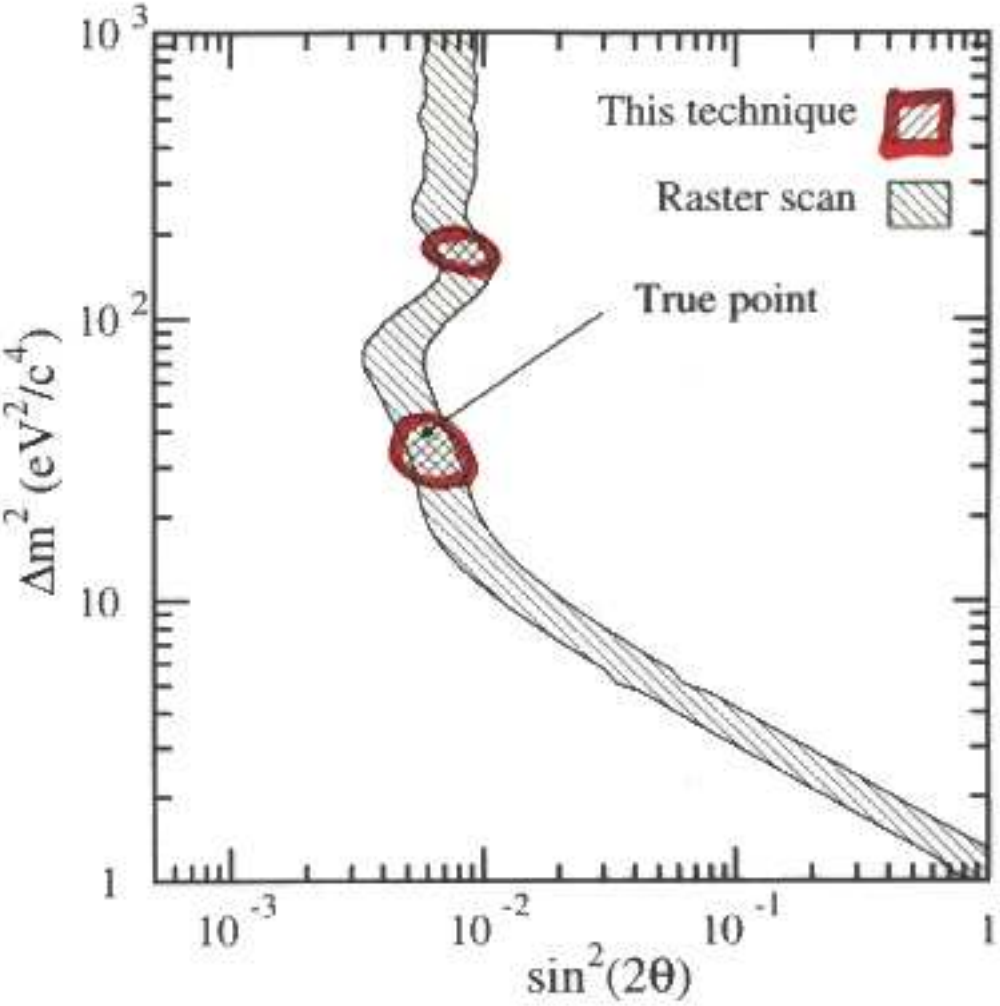


FIG. 12. Calculation of the confidence regions for an example of the toy model in which $\Delta m^2 = 40$ $(\mathrm{eV}/c^2)^2$ and $\sin^2(2\theta) = 0.006$, as evaluated by the proposed technique and the Raster Scan.

i.e. FELDMAN - COUSINS IS MUCH BETTER THAN RASTER SCAN

(cf $B-\bar{B}$ OSCILLATIONS)

SENSITIVITY
(Indep of actual data)

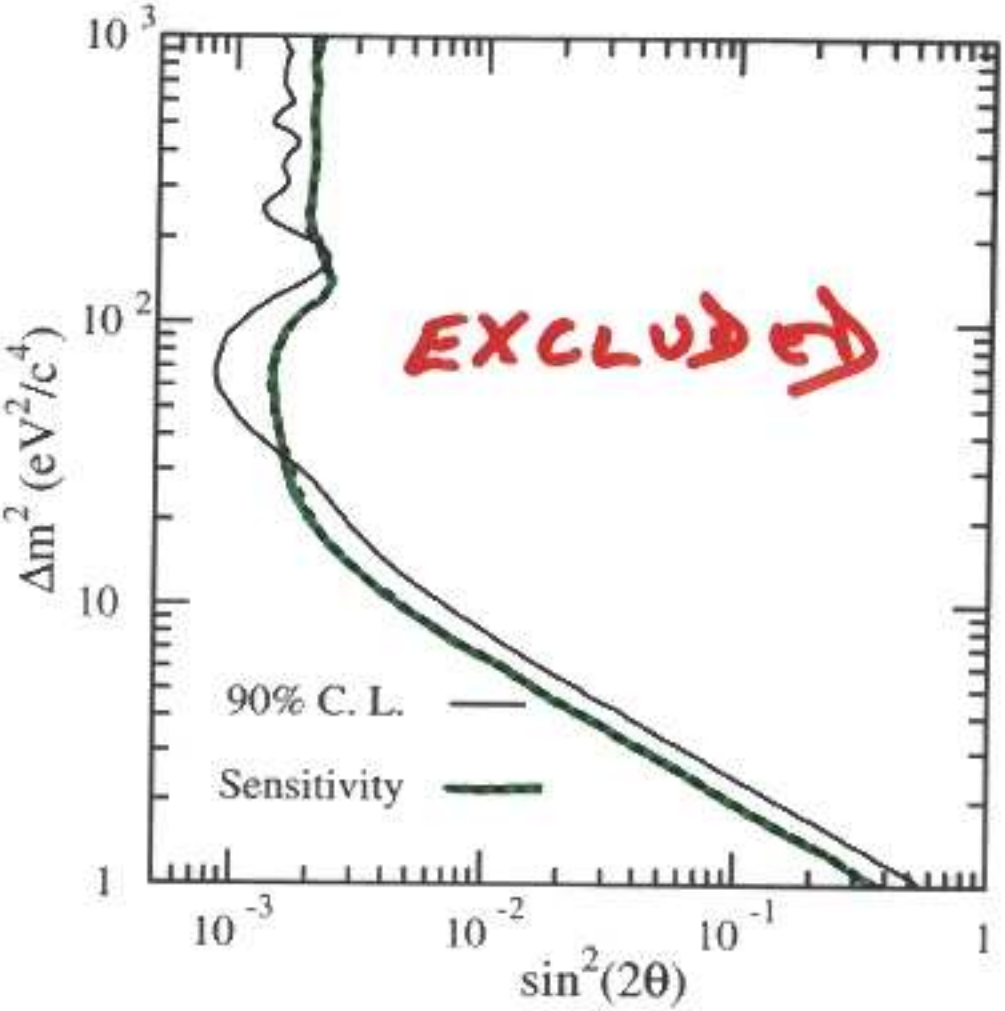
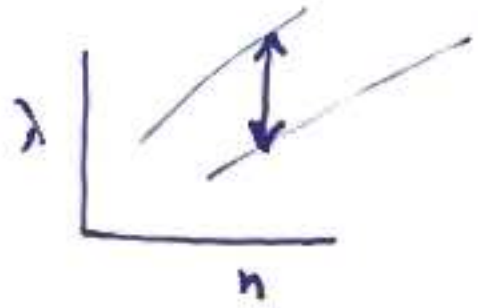
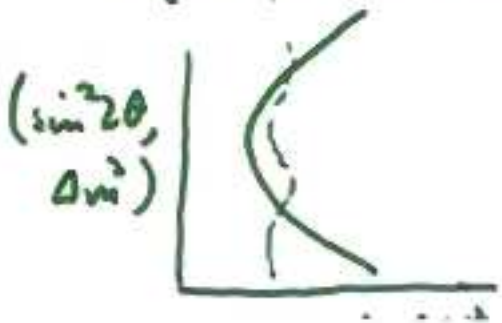


FIG. 15. Comparision of the confidence region for an example of the toy model in which $\sin^2(2\theta) = 0$ and the sensitivity of the experiment, as defined in the text.

	1-D Poisson	ν oscillations
PARAM	λ	$\sin^2 2\theta$, Δm^2
DATA	n	n_i (E_ν)
ORDERING RULE	$\ln[\mathcal{L}$ RATIO$]$	$\ln[\mathcal{L}_{\text{RATIO}}] \sim \Delta\chi^2$
HOW MANY?	At each λ, $\sum P(n\|\lambda) = 0.9$	At each $(\sin^2 2\theta, \Delta m^2)$, include first 90% of $\Delta\chi^2$
ACCEPTANCE REGION	USE OBSERVED n	USE OBSERVED $n_i(E_\nu)$ $\Rightarrow$ $\Delta\chi^2(\sin^2 2\theta, \Delta m^2)$
	$\Rightarrow$ λ RANGE	$\Rightarrow$ $(\sin^2 2\theta, \Delta m^2)$ REGION
	[REGION WHERE $\lambda \approx n$]	[REGION WHERE $n_i \approx$ prediction for $(\sin^2 2\theta, \Delta m^2)$]





SYSTEMATICS

For example

$$N_{events} = \sigma \, LA + b$$

↑ Observed ↑ $N \pm \sqrt{N}$ for statistical errors

↑ Physics parameter

↑ ↑ we need to know these, probably from other measurements (and/or theory)

Uncertainties → error in σ

Some are arguably statistical errors

$$LA = LA_o \pm \sigma_{LA}$$

$$b = b_o \pm \sigma_b$$

Shift Central Value

Bayesian

Frequentist

Mixed

Profile Likelihood

$$N_{events} = \sigma \, LA + b$$

Simplest Method

Evaluate σ_0 using LA_0 and b_0

Move nuisance parameters (one at a time) by their errors → $\delta\sigma_{LA}$ & $\delta\sigma_b$

If nuisance parameters are uncorrelated

Combine these contributions in quadrature

→ total systematic

PROFILE $\mathcal{L}$

Rolke, Lopez, Conrad + James

"Limits & Confidence Intervals in the Presence of Nuisance Parameters"

$$\mathcal{L}_p(\mu|\text{data}) = \mathcal{L}(\mu, b_{best}|\text{data})$$

$$\Delta \ln \mathcal{L}_p = 0.5$$

Coverage much smoother (as fn of μ) than for standard Bayesian without nuisance parameters

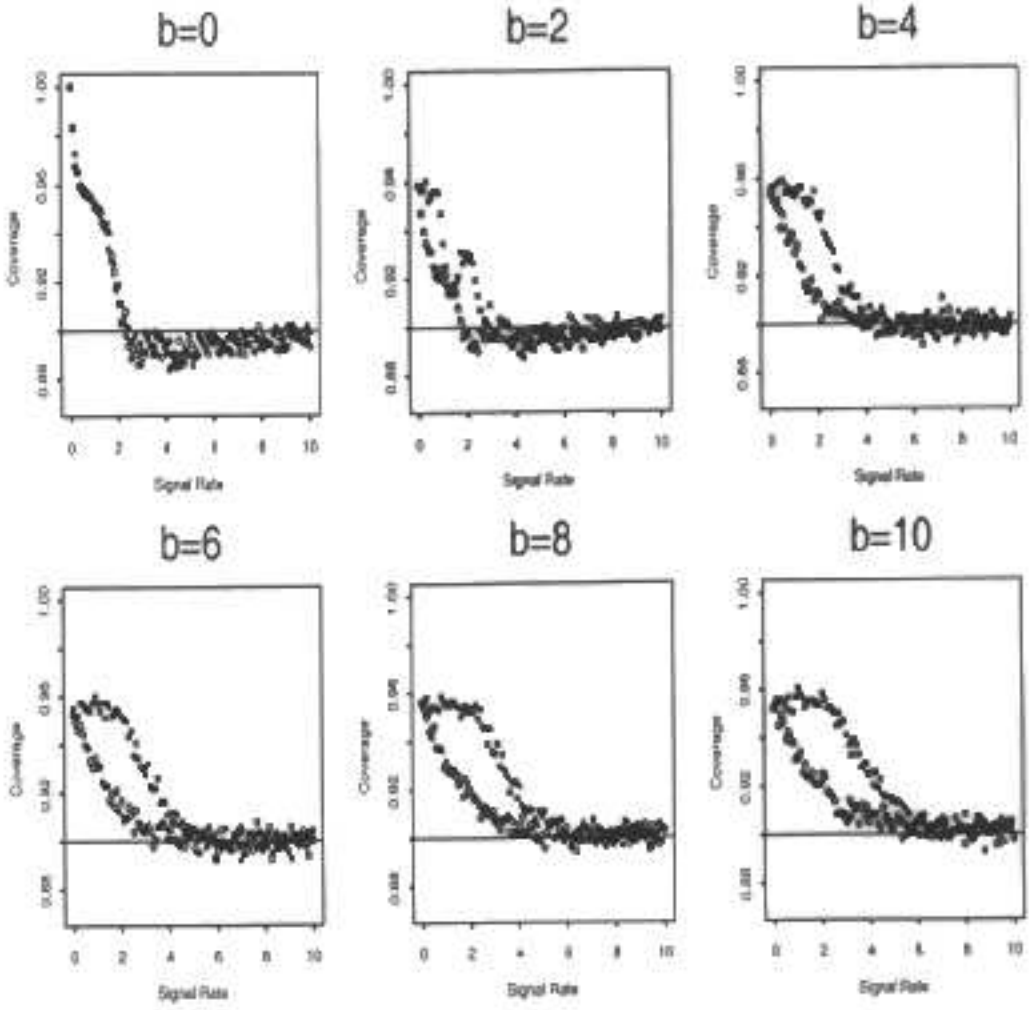


FIG. 3: 90% coverage graphs when the signal and the background are modeled as Poisson and the efficiency is modeled as a Binomial. We have $\tau = 3.5$, $\epsilon = 0.85$ and $m = 100$. The empty circles show the coverage using the unbounded likelihood method and the solid squares show the coverage using the bounded likelihood method.

Rolke et al

Profile L

Bayesian

Without systematics

$$p(\sigma; N) \propto p(N; \sigma)\Pi(\sigma)$$

↑ prior

With systematics

$$p(\sigma, LA, b; N) \propto p(N; \sigma, LA, b)\Pi(\sigma, LA, b)$$

↑

$$\sim \Pi_1(\sigma)\Pi_2(LA)\Pi_3(b)$$

Then integrate over LA and b

$$p(\sigma; N) = \iint p(\sigma, LA, b; N)\, dLA\, db$$

$$p(\sigma; N) = \iint p(\sigma, LA, b; N)\, dLA\ db$$

If $\Pi_1(\sigma)$ = constant and $\Pi_2(\mathrm{LA})$ = truncated Gaussian TROUBLE!

Upper limit on σ from $\int p(\sigma; N)\, \mathrm{d}\sigma$

Significance from likelihood ratio for $\sigma = 0$ and $\sigma_{\max}$

BAYES 90% UPPER LIMITS

	$\epsilon = 1.0 \pm 0.1$		$\epsilon = 1$ exactly	
Bgd	0	3	0	3
n_{obs}				
0	2.35 indep of b		2.30 indep of b	
1	3.99	2.90	3.89	2.84
2	5.47	3.60	5.32	3.52
3	6.87	4.46	6.68	4.36
4	8.24	5.48	7.99	5.34
⋮	⋮	⋮	⋮	⋮
20	28.3	25.04	27.05	24.04

Less than 10% bigger than for $\epsilon = 1$ exactly

$\Delta = 0$ for $b = 0$

$\Delta = 3$ for large b

$\sim n + \kappa\sqrt{n}$

JOEL HEINRICH et al (CDF St. Ctteee)

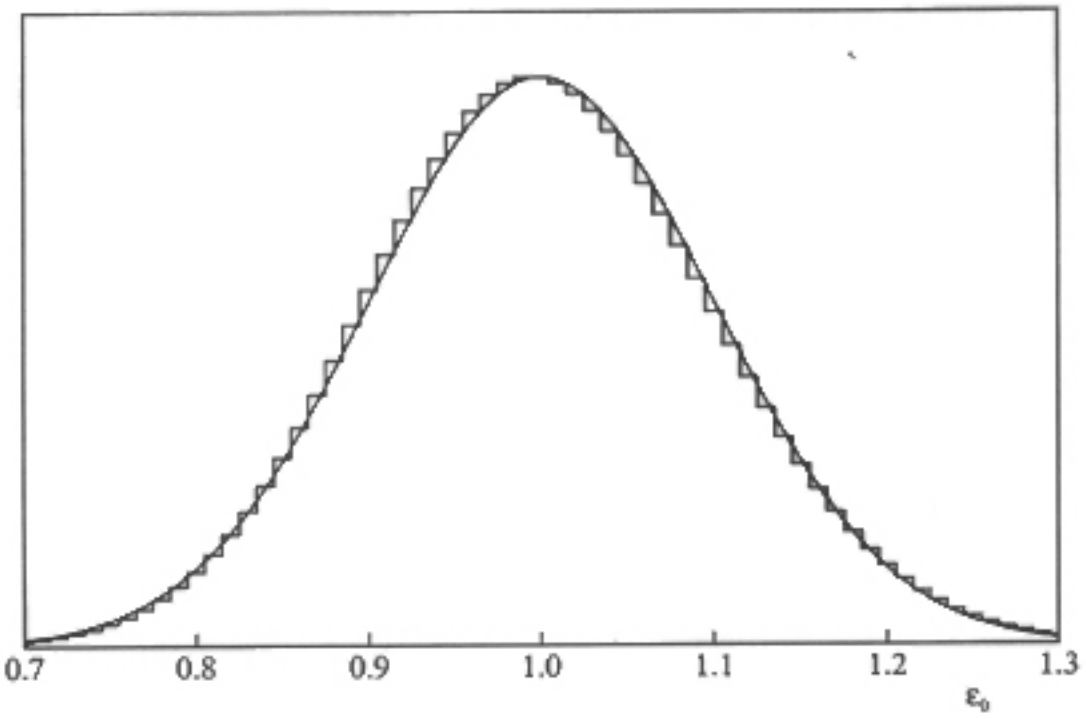


Figure 2: Comparison of our discrete probability for ϵ_0 (shown as a histogram, see eqn (11)) and Gaussian (continuous curve) for the case $\epsilon = 1 \pm 0.1$.

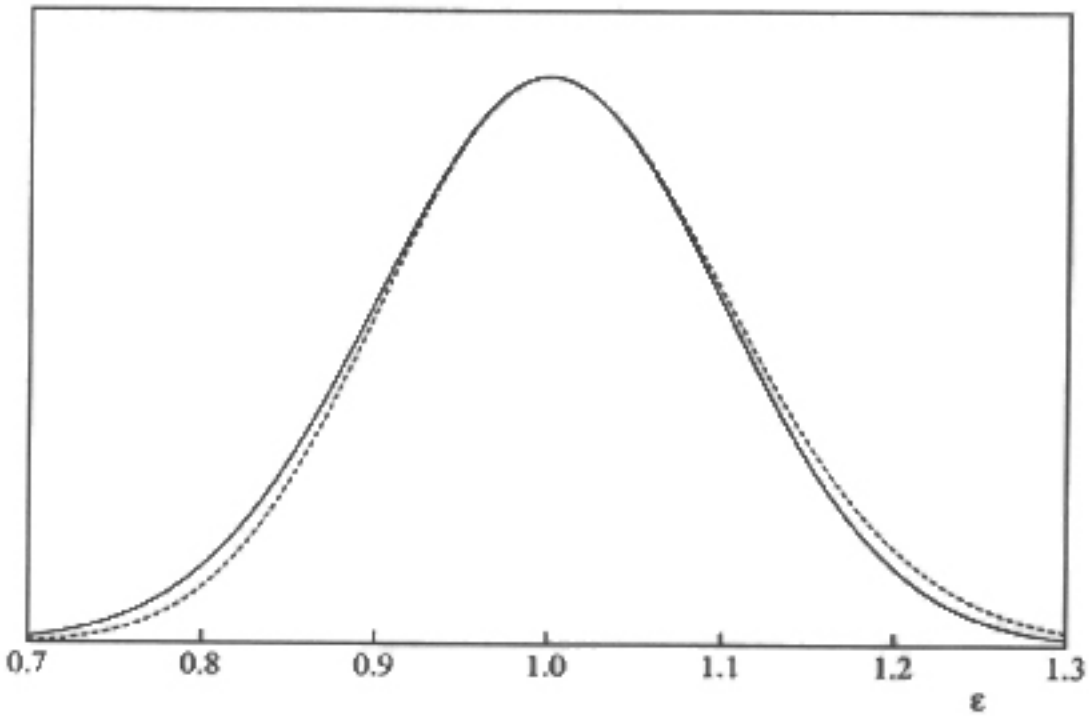


Figure 3: Comparison of our likelihood (dashed, see eqn (12)) and Gaussian (solid) for the case $\epsilon = 1 \pm 0.1$.

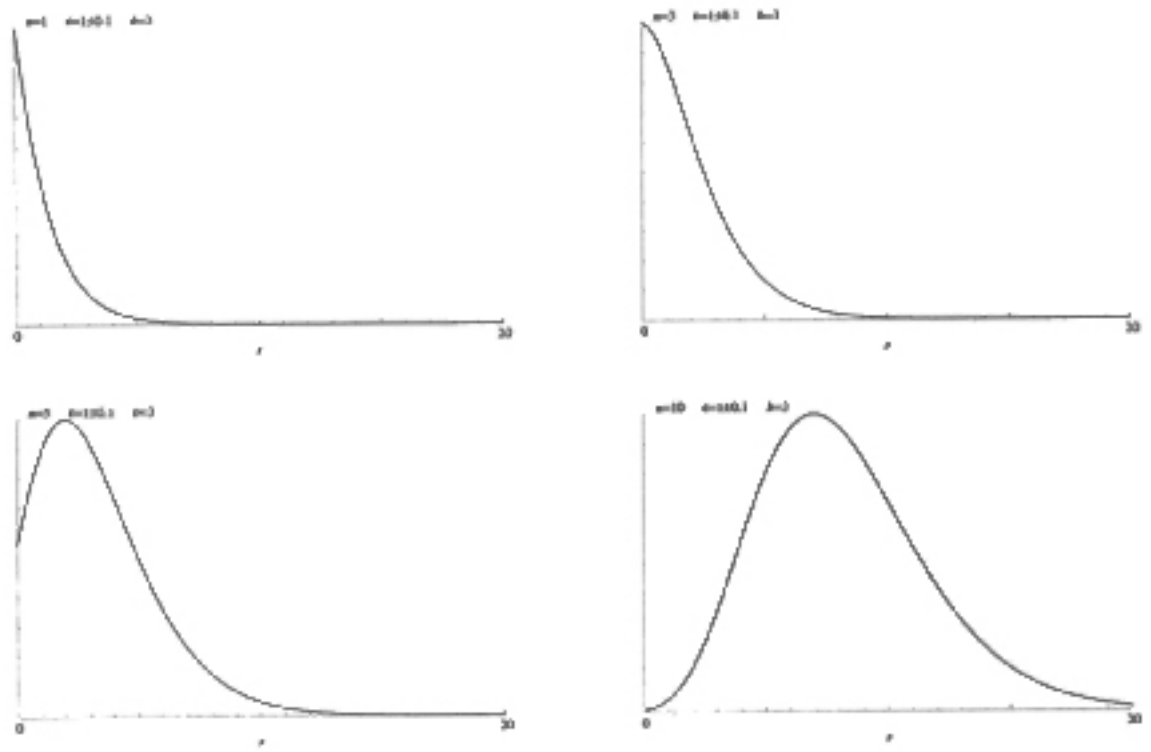


Figure 4: Posterior densities $p(s|n,b)$ vs s for $n = 1$, 3, 5, 10. In each case, $b = 3$ and $\epsilon = 1 \pm 0.1$ (i.e. $\kappa = 100$ and m=99).

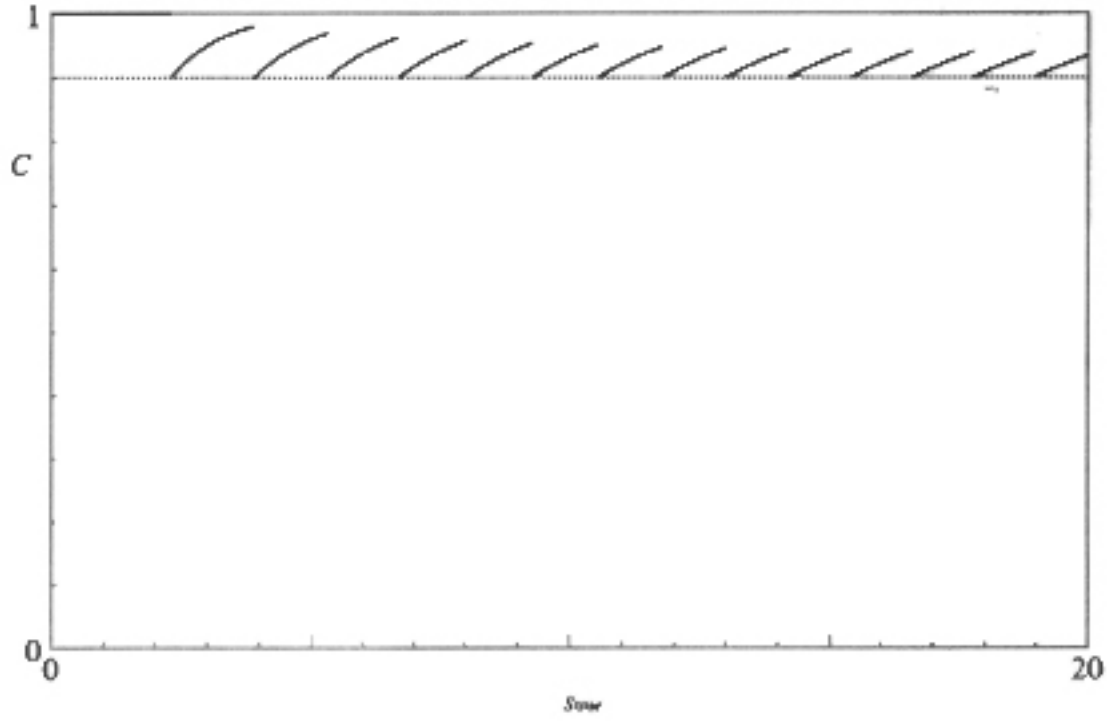


Figure 1: Coverage as a function of the true signal rate s for Bayes 90% limits, for the simple case of no background and no uncertainty on $\epsilon = 1$.

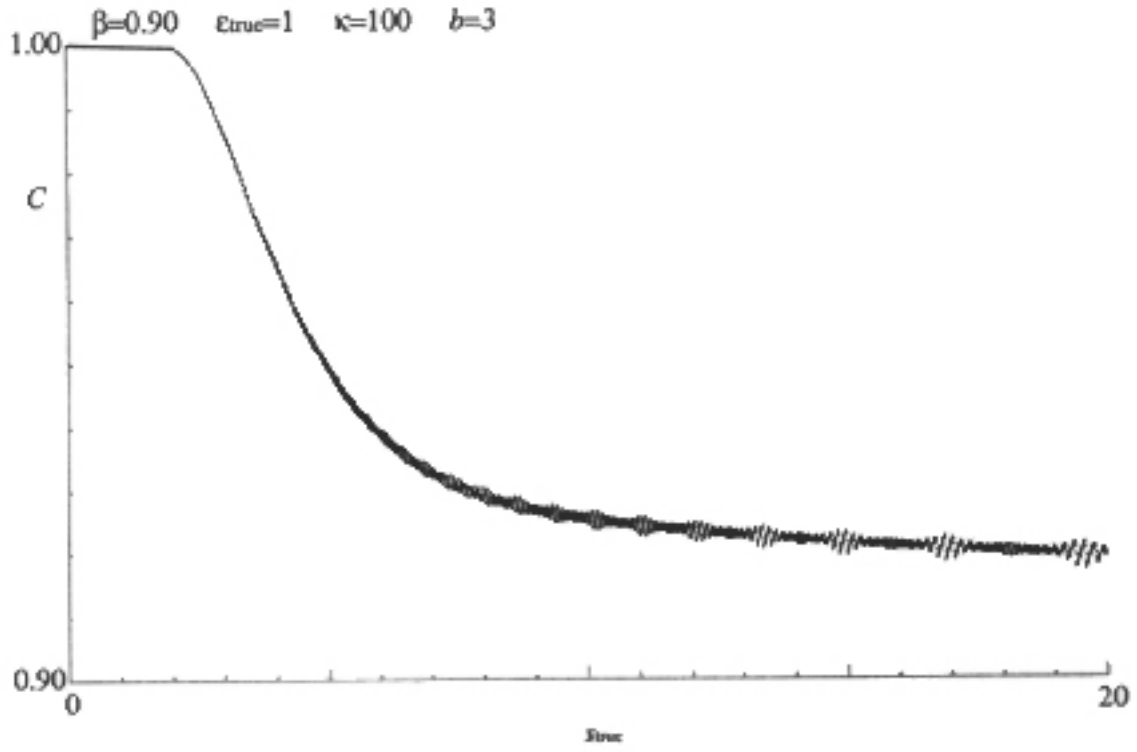


Figure 5: Coverage of 90% upper limits as a function of s_{true} for $\epsilon_{\text{true}} = 1$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

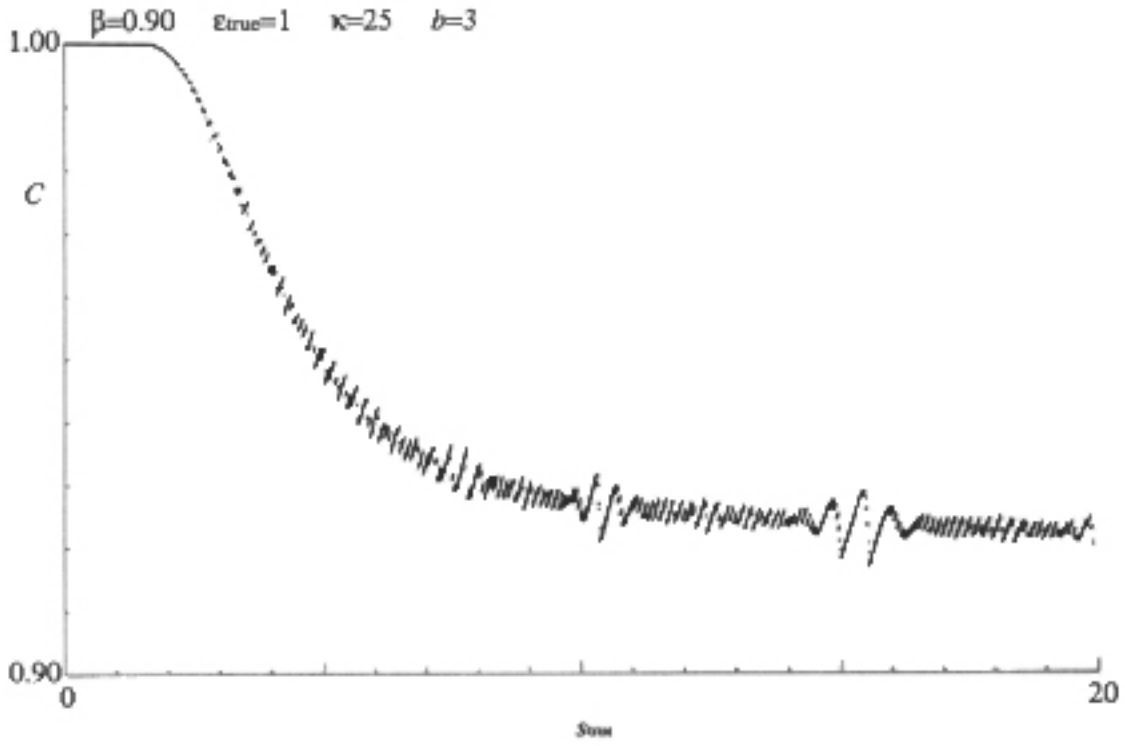


Figure 6: Coverage of 90% upper limits as a function of s_{true} for $\epsilon_{\text{true}} = 1$, nominal 20% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

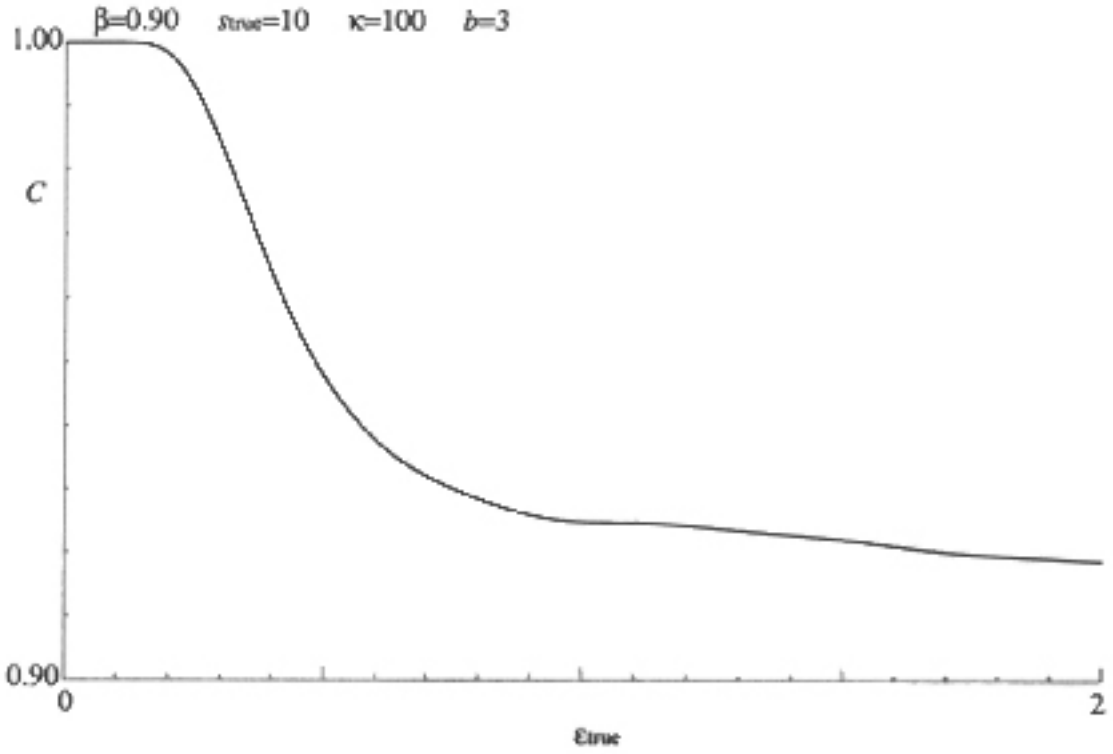


Figure 7: Coverage of 90% upper limits as a function of ϵ_{true} for $s_{\text{true}} = 10$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

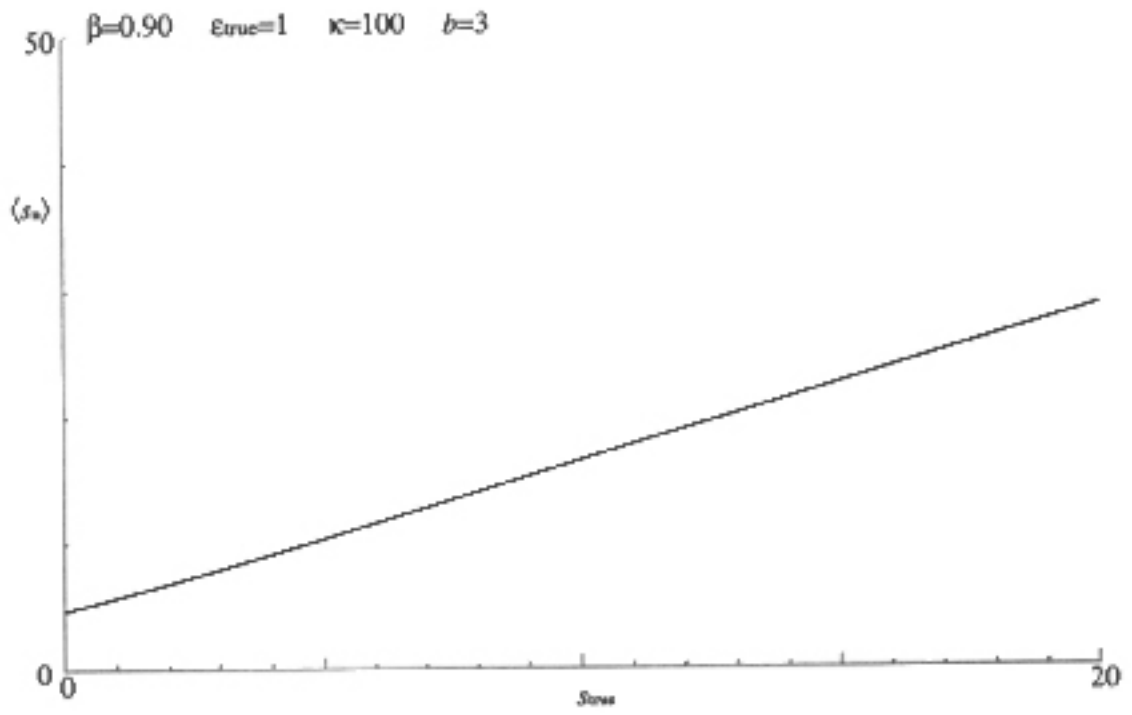


Figure 8: Sensitivity of 90% upper limits as a function of s_{true} for $\epsilon_{\text{true}} = 1$, nominal 10% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

Frequentist

Full Method

Imagine just 2 parameters σ and LA

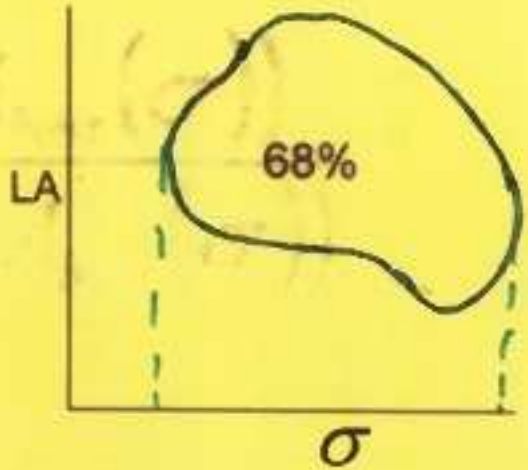
and 2 measurements N and M

↑ Physics ↑ Nuisance

Do Neyman construction in 4-D

Use observed **N** and **M**, to give

Confidence Region



Then project onto σ axis

This results in OVERCOVERAGE

Aim to get better shaped region, by suitable choice of ordering rule

Example: Profile likelihood ordering

$$\frac{\mathrm{L}(N_0 M_0; \sigma, LA_{best}(\sigma))}{\mathrm{L}(N_0 M_0; \sigma_{best}, LA_{best}(\sigma))}$$

Full frequentist method hard to apply in several dimensions

Used in ≤ 3 parameters

For example: Neutrino oscillations (CHOOZ)

$$\sin^2 2\theta \,, \Delta m^2$$

Normalisation of data

Use approximate frequentist methods that reduce dimensions to just physics parameters

e.g. Profile pdf

i.e. $pdf_{profile}(N;\sigma) = pdf(N, M_0; \sigma, LA_{best})$

Contrast Bayes marginalisation

Distinguish "profile ordering"

Properties being studied by Giovanni Punzi

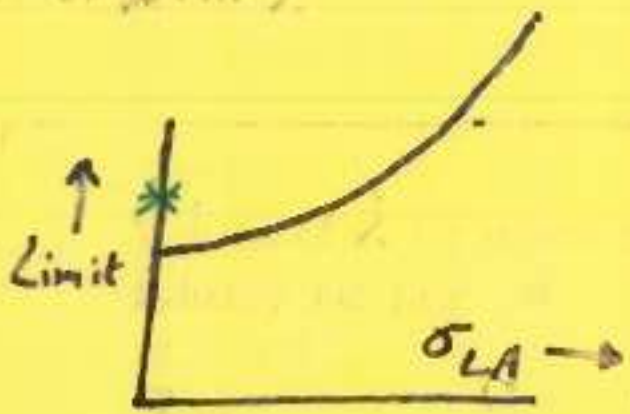
Method: Mixed Frequentist - Bayesian

Bayesian for nuisance parameters and

Frequentist to extract range

Philosophical/aesthetic problems?

Highland and Cousins



(Motivation was paradoxical behavior of Poisson limit when LA not known exactly)

Coverage studied by Tegenfeldt + Conrad

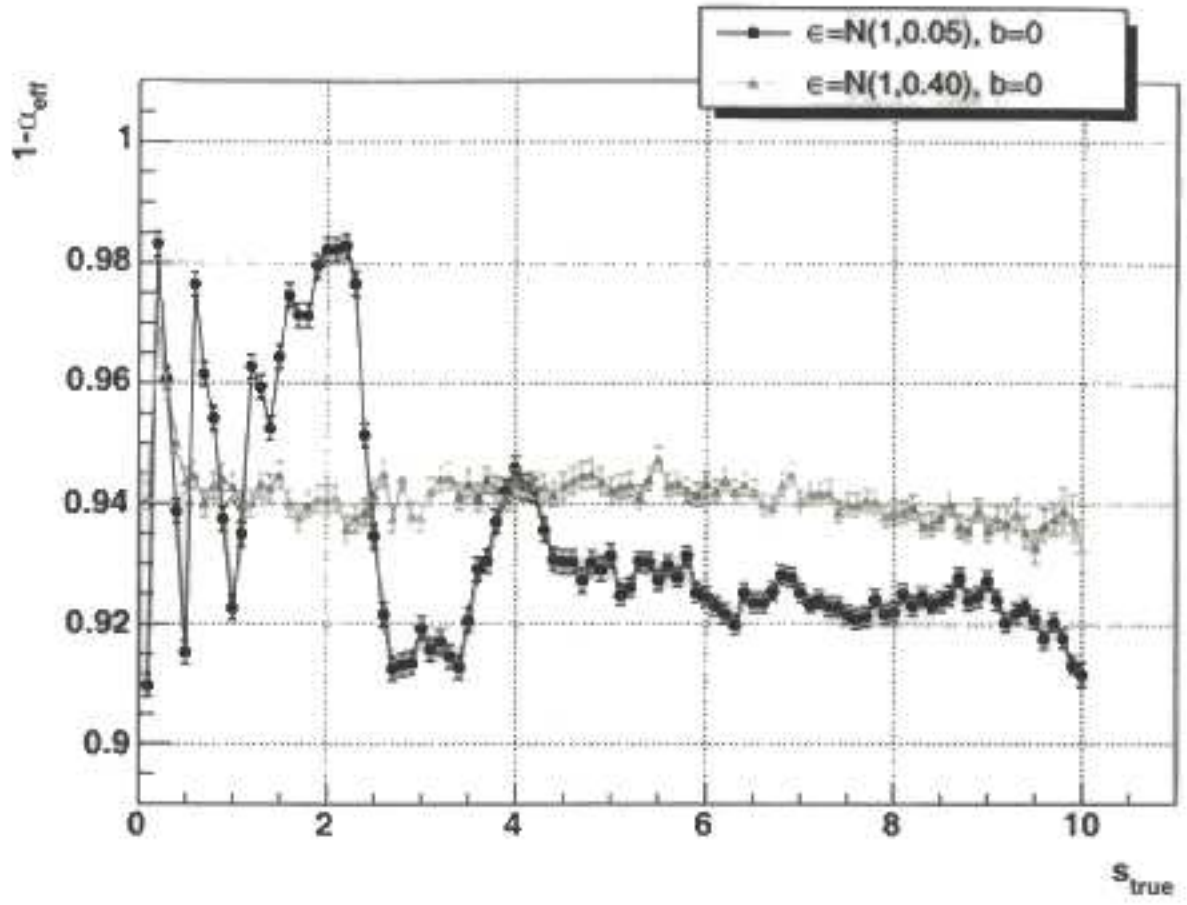


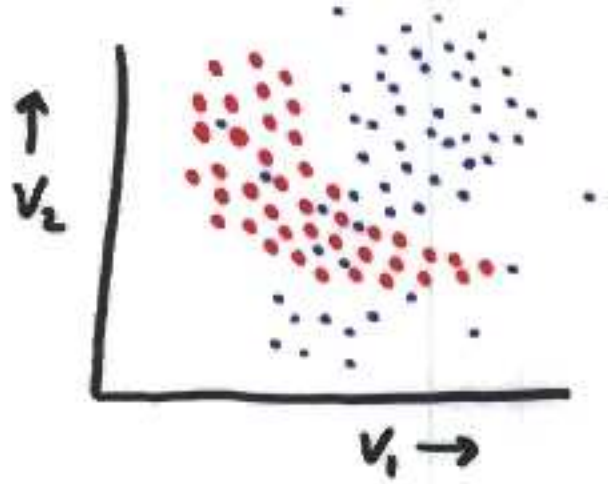
Fig. 1. Calculated coverage as function of signal hypothesis. Two case are shown: 5 % and 40 % Gaussian uncertainties in the signal efficiency. The nominal coverage was 90%.

Tegenfeldt + Conrad

Feldman - Cousins + Bayes

MULTIVARIATE ANALYSIS

AIM TO SEPARATE SIGNAL FROM BACKGROUND



NEYMAN - PEARSON THEOREM

1) IMAGINE ALL POSSIBLE CONTOURS THAT SELECT SIGNAL WITH α% EFFICIENCY
(LOSS = ERROR OF 1ST KIND)

2) BEST IS ONE CONTAINING MINIMAL AMOUNT OF BG (CONTAM = ERROR OF 2ND KIND)

EQUIVALENT TO ORDERING DATA BY

$$\mathcal{L}\text{-RATIO} = \frac{\mathcal{L}_S(v_1, v_2 \dots)}{\mathcal{L}_B(v_1, v_2 \dots)}$$

IF VARIABLES INDEPENDENT,

$$\rightarrow \frac{\mathcal{L}_S(v_1)}{\mathcal{L}_B(v_1)} \times \frac{\mathcal{L}_S(v_2)}{\mathcal{L}_B(v_2)} \times \dots$$

PROBLEM: DON'T KNOW $\mathcal{L}$-RATIO EXACTLY BECAUSE:—

1) GENERATED BY M.C. WITH FINITE STATISTICS

2) UNCERTAIN PARAMETERS (NUISANCE PARAMS, SYSTEMATICS)

3) NEGLECTED SOURCES OF BKD

4) HARD TO IMPLEMENT N-P IN MANY DIMENSIONS

METHODS: CUTS

FISHER DISCRIMINANT

PRINCIPAL COMPONENT ANALYSIS

INDEPENDENT COMP. ANALYSIS

BOOSTED TREE METHODS

KERNEL DENSITY ESTIMATION

NEURAL NETS ★ ★

SUPPORT VECTOR MACHINES

⋮

USEFUL REFERENCES

H. PROSPER : 'MULTIVARIATE ANALYSIS' (DURHAM)

J. FRIEDMAN : 'PREDICTIVE MACHINE LEARNING' (PHYSTAT 2003)

R. BOCK : 'MULTIDIM. EVENT CLASSIFICATION FOR γ RAY SHOWERS (DURHAM)

FNAL AAG http://projects.fnal.gov/run2aag/

S. TOWERS : i) PROB DENSITY ESTIMATION.
ii) REDUCE NUMBER OF VARIABLES
(BOTH AT DURHAM)

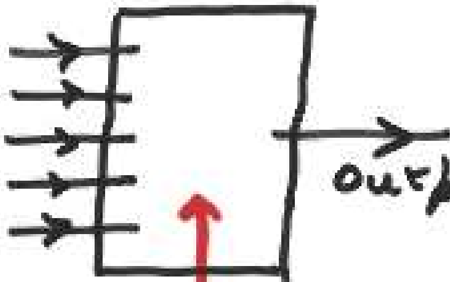
A VAICIULIS : SUPPORT VECTOR MACHINES (DURHAM)

NEURAL NETWORKS

TYPICAL APPLICATION:

CLASSIFY EVENTS AS { SIGNAL / BACKGROUND

Inputs



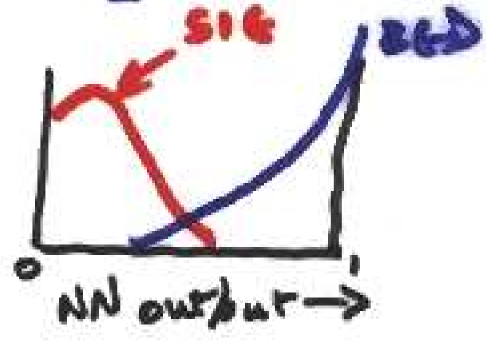
{ ~1 FOR SIGNAL / ~0 FOR BGD

Adjustable params (WEIGHTS + THRESHOLDS)

1) LEARNING PROCESS:

INPUT = { KNOWN SIGNAL / KNOWN BGD ← M.C?

ADJUST PARAMS ⇒ 'BEST' OUTPUT



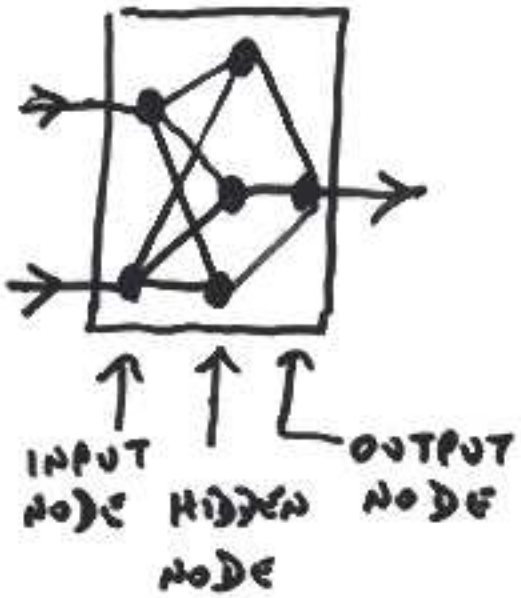
2) TESTING PROCESS

MAKE SURE NO "OVERTRAINING"

3) USE TRAINED NET ON ACTUAL DATA.

CLASSIFY AS SIGNAL IF NN OUTPUT > Cut

How does it work?



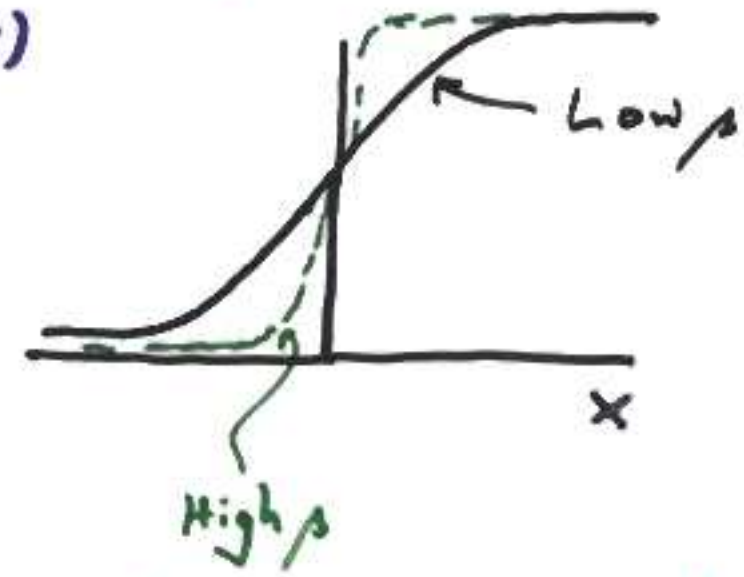
For each node

$$\text{Output} = F\left[\sum \text{Input}_i \times W_i + T\right]$$

(W and T: Params)

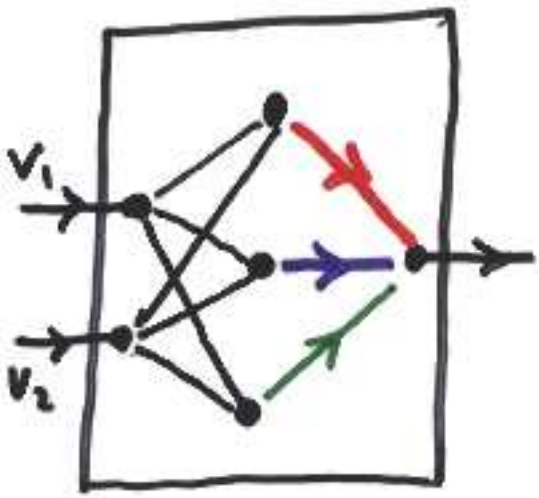
Typical $F(x)$

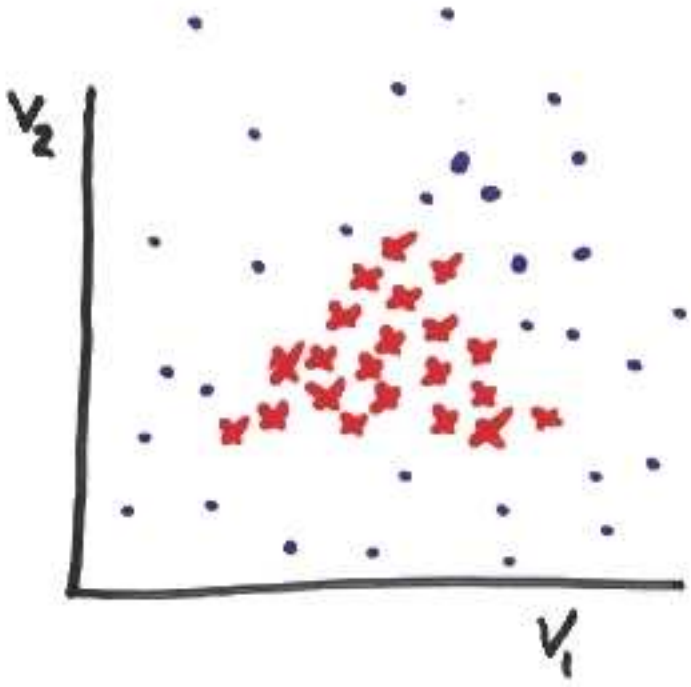
$$\frac{1}{1+e^{-\beta x}}$$

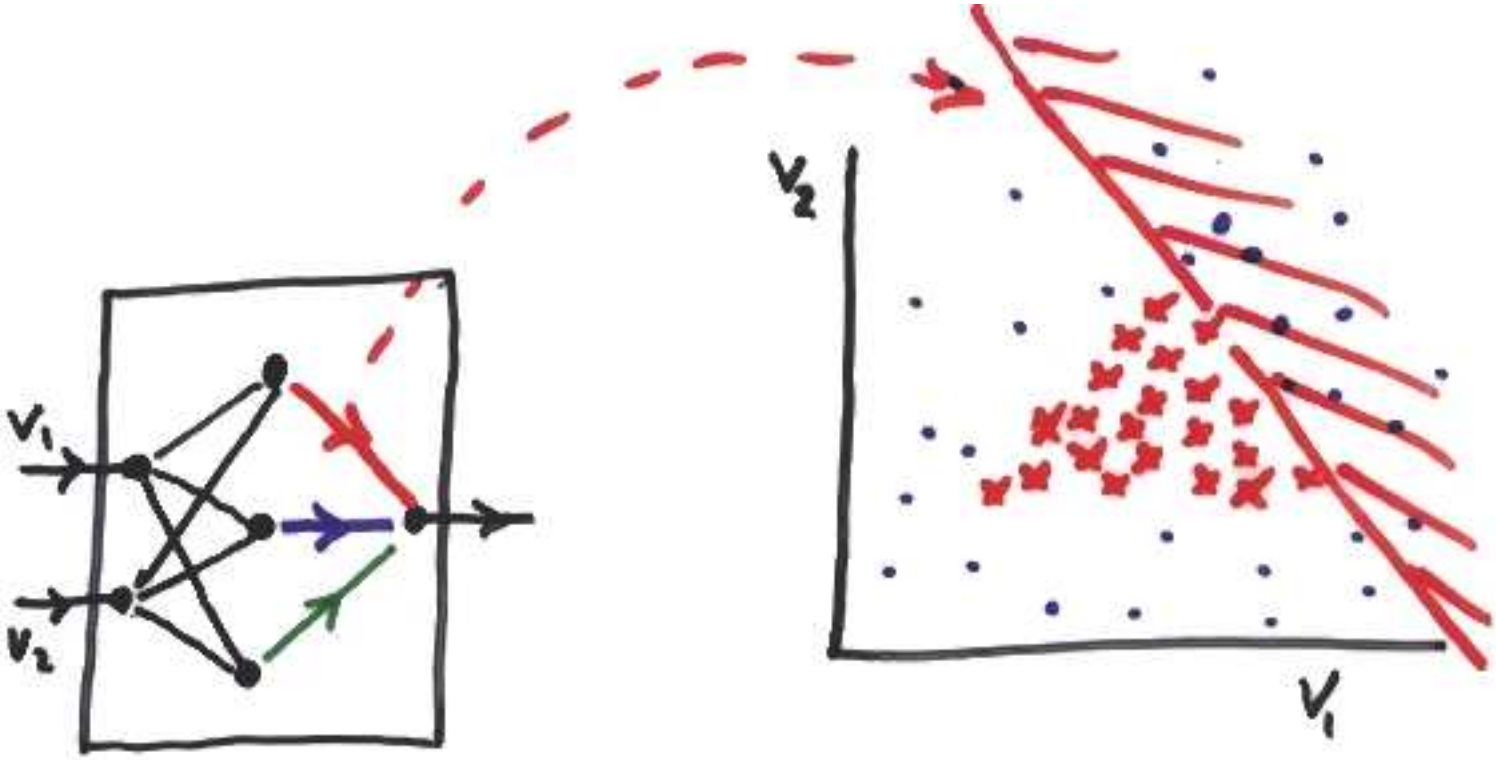


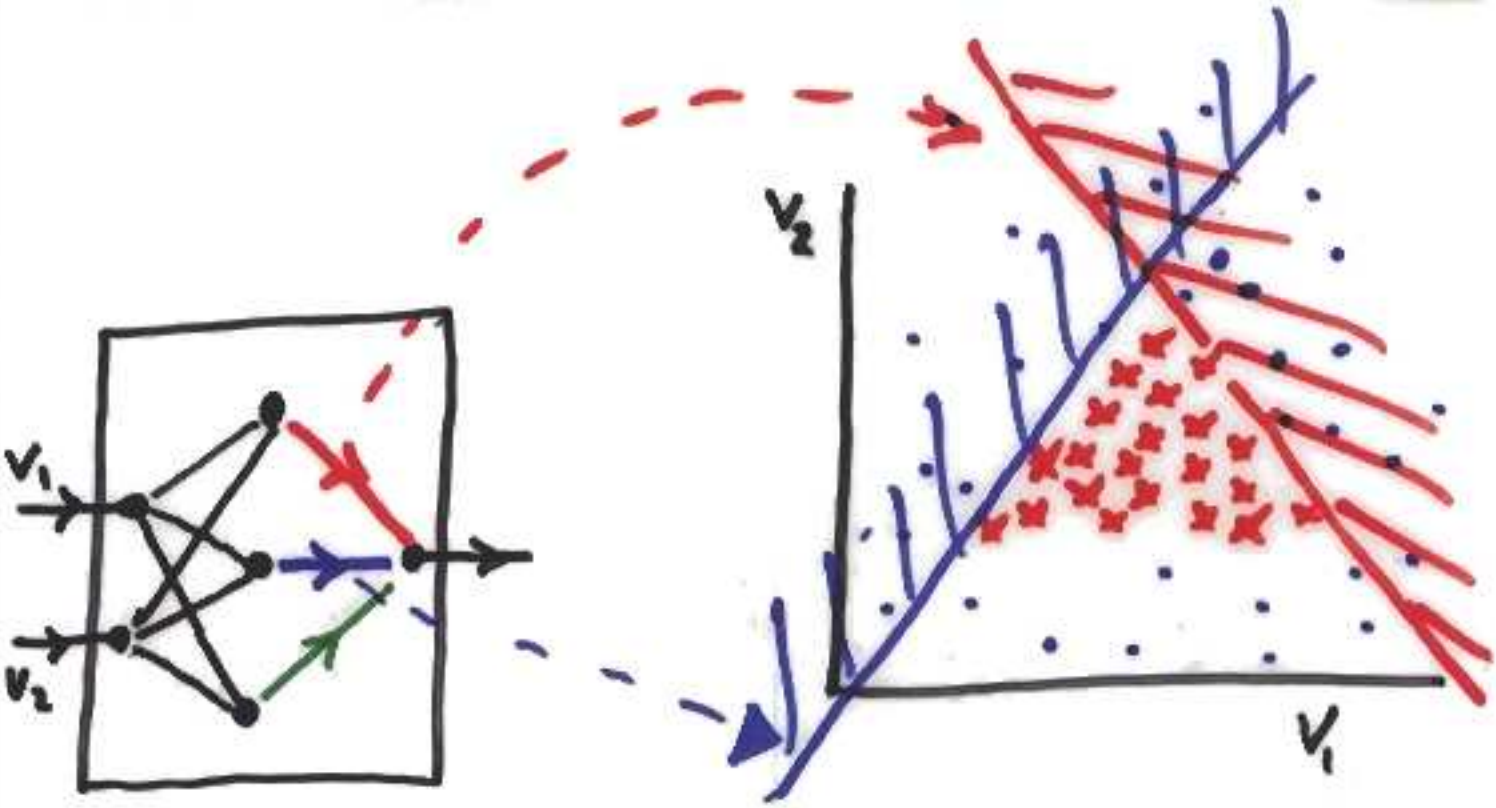
⇒ Output of node is "ON" if $\sum I_i W_i + T > 0$

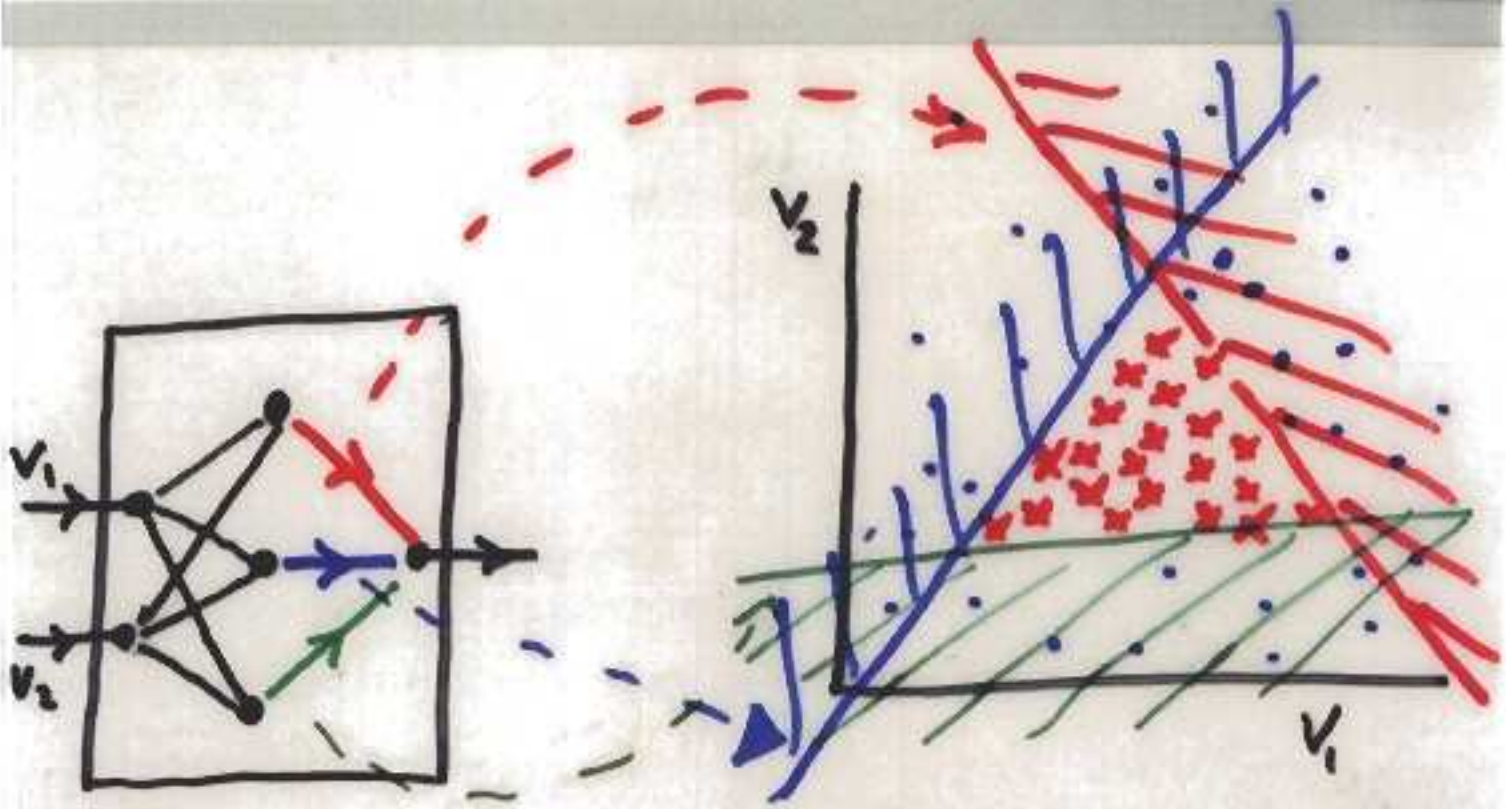
This is "hyper-plane" in I. space


V_1
V_2


V_2
V_1


V1
V2
V2
V1


V_1
V_2
V_2
V_1



$$\text{Output} = F\left[0.4H_1 + 0.4H_2 + 0.4H_3 - 1.0\right]$$

Output = "ON" only if H_1 H_2 H_3 all are "ON"

N.B.

1) Complexity of final region depends on number of hidden nodes

2) Finite $\beta \Rightarrow$ rounded edges for selected region.

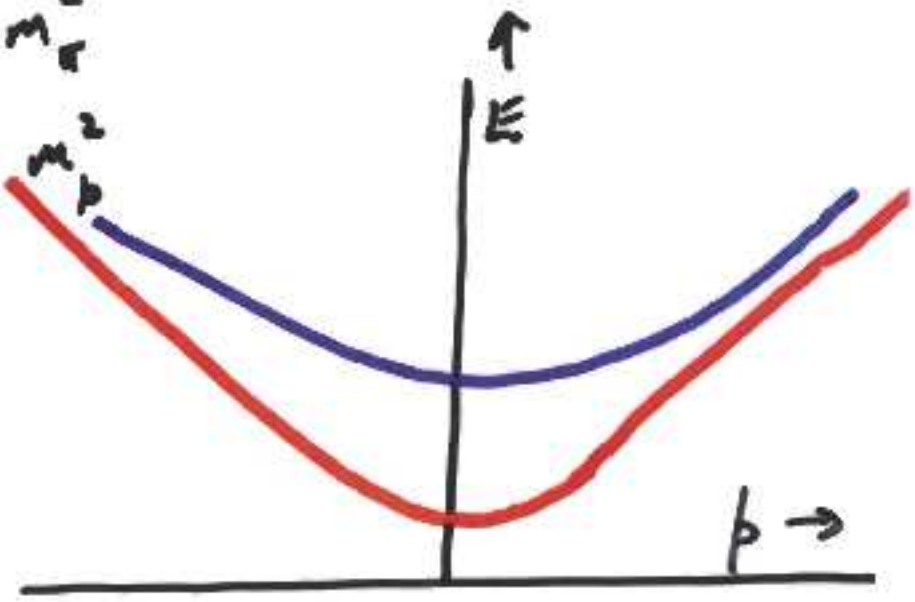
[Output contour lines in V_1-V_2 plane]

TOY EXAMPLE

Try to separate $\left\{ \begin{matrix} \pi \\ p \end{matrix} \right\}$ using p or E

π: $E^2 = p^2 + m_\pi^2$

p: $E^2 = p^2 + m_p^2$



Easy : $p = 0 \rightarrow 2$ GeV/c

Harder : $p = -4 \rightarrow +4$ GeV/c

Hardest $\left\{ \begin{matrix} p_x \\ p_y \\ p_z \end{matrix} \right\} = -4 \rightarrow +4$ GeV/c

More realistic : Add scatter of data about curves

Is NN better than simple cuts?

In principle, NO

[Can cut on complicated variable
e.g. NN output]

In practice, YES (usually)

But better NN performance

⇒ motivation to improve cuts.

PHYSICS EXAMPLE

Separate $e^+e^- \to c\bar{c}$.

from ; $b\bar{b}$, $q\bar{q}$, W^+W^-, ZZ

at LEP

Input variables : "Lifetime"

Track rapidities

Secondary vertex mass

quality

etc.

ISSUES: PRE N-N. CUTS

MISSING VARIABLES

WHERE TO GET TRAINING/TESTING EVENTS

HOW MANY NN's.

HOW MANY INPUT VARIABLES

HOW MANY HIDDEN NODES/LAYERS

SINGLE OUTPUT OR SEVERAL

RATIO OF $c\bar{c}$, $b\bar{b}$, $q\bar{q}$, ... TRAINING EVENTS

SYSTEMATICS [USE DIFFERENT SETS OF TESTING EVENTS]

STABILITY W.R.T. NN CUT

NN SUMMARY

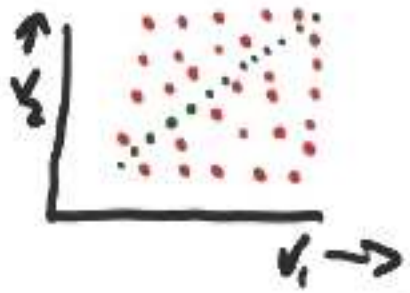
ADVANTAGES:

VERY FLEXIBLE

CORRELATIONS O.K.

TUNABLE CUT

e.g. minimum σ



DISADVANTAGES

TRAINING TAKES TIME

TENDENCY TO INCLUDE TOO MANY VARIABLES

TREAT AS BLACK BOX

OVER LAST FEW YEARS, CHANGE IN ATTITUDE FROM:

CONVINCE COLLEAGUES NN IS SENSIBLE

TO

"WHY DON'T YOU USE NN?"

BLUE

Best Linear Unbiassed Estimate

$x_i \pm \sigma_i$, possibly correlated

$$\hat{x} = \sum \alpha_i x_i \qquad \sum \alpha_i = 1$$

$$\sigma^2_{\hat{x}} = \text{minimum}$$

LL , Duncan Gibaut & Peter Clifford
NIM A270 (1988) 110

For contemplation:

Given x_1 and x_2 (+ error matrix),

Can $\hat{x}$ lie outside range $x_1 \to x_2$?