χ^2 and Goodness of Fit

Louis Lyons and Lorenzo Moneta
Imperial College & Oxford
CERN

CERN Academic Training Course
Nov 2016

Least squares best fit

- What is σ?
- Resume of straight line
- Correlated errors

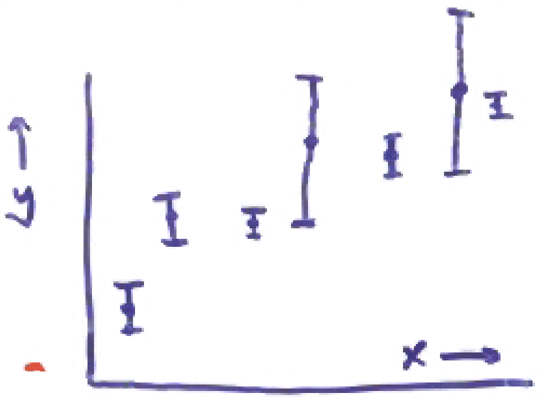
Goodness of fit with χ^2

- Number of Degrees of Freedom
- Other G of F methods
- Errors of first and second kind
- Combinations

THE paradox

Least Squares Straight Line Fitting

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



1) Does it fit straight line?

(Goodness of Fit)

2) What are gradient and intercept?

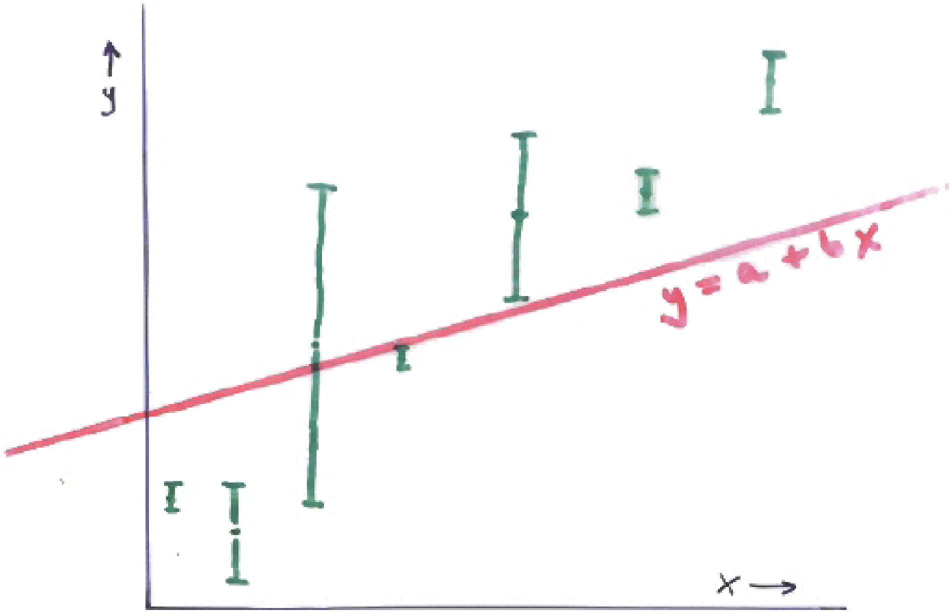
(Parameter Determination)

Do 2) first

N.B.1 Can be used for non "a+bx"

e.g. $a + b/x + c/x^2$

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



Criterion:

$a + bx_i$; Vert devn

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

An error for each pt.

If theory and data OK:

$y^{th} \sim y^{obs}$ → S small

Minimise S → best line

Value of S_{min} → how good fit is

Which σ should we use?

Which σ?	Exptl σ	Theory σ
Name	Neyman	Pearson
Ease of algebra	Easier, so this version is used more	
If Th = 0.01, Exp = 1	Contributes 1 to S	Contributes 98 to S More plausible
S ~ χ^2 ?		More plausible
$S = (\hat{a} - a_1)^2/\sigma^2 + (\hat{a} - a_2)^2/\sigma^2$	Biassed down because smaller $a_i \rightarrow$ smaller σ	Biassed up because larger $\hat{a} \rightarrow$ larger σ

(For $\hat{a} \sim a_i$, and both much larger than σ_i, 2 methods are very similar)

Straight Line Fit

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

(Fixed σ_i)

i) "Draw" lots of lines ⇒ S for each

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

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

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

2 SIM. EQNS FOR 2 UNKNOWNS (a & b)

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

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

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

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

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

Correlated intercept and gradient?

2 * Inverse covariance matrix =

$$\begin{bmatrix} \frac{\partial^2 S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial a \partial b} & \frac{\partial^2 S}{\partial b^2} \end{bmatrix} = \begin{bmatrix} \Sigma 1/\sigma_i^2 & \Sigma x_i/\sigma_i^2 \\ \Sigma x_i/\sigma_i^2 & \Sigma x_i^2/\sigma_i^2 \end{bmatrix}$$

Invert→ Covariance matrix

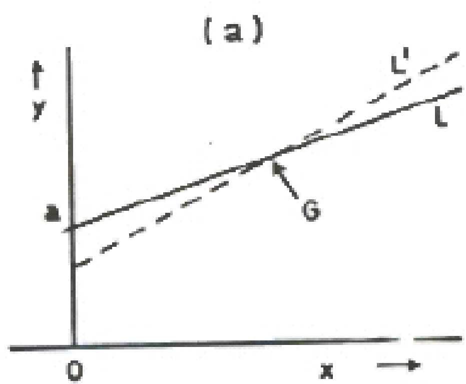
Covariance ~ $-\Sigma x_i/\sigma_i^2$ = [x]

If measure intercept at weighted c. of g. of x for data points, cov = 0

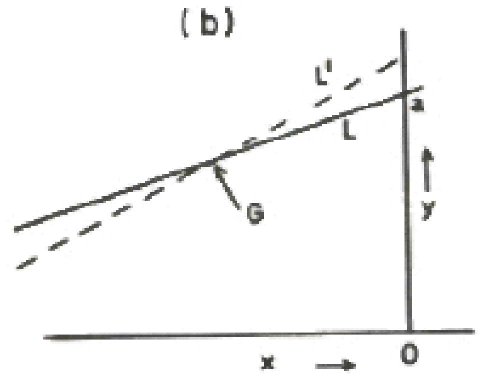
i.e. gradient and intercept there are uncorrelated

So track params are usually specified at centre of track.

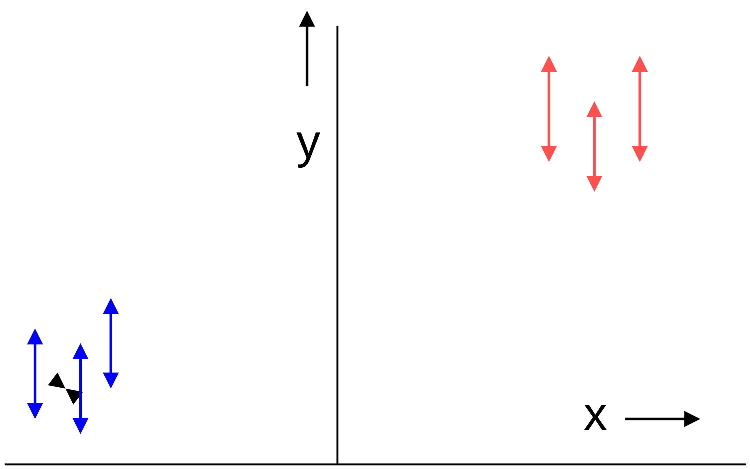
Covariance(a,b) ~ -<x>

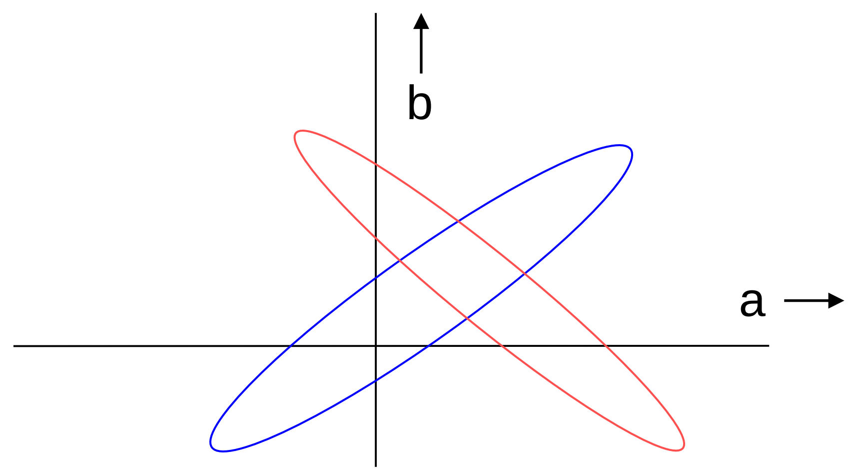


<x> positive

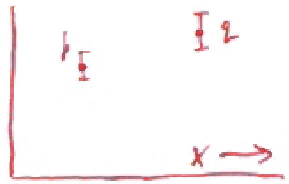


<x> negative





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

$$\left.\begin{aligned} p &= r\cos\theta - s\sin\theta \\ q &= r\sin\theta + s\cos\theta \end{aligned}\right\}$$ NOT ROTN in x-y SPACE

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

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

Inv. err matrix element →

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

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

Reduces to standard formula in absence of correlns

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

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

Comments on Least Squares method

1) Need to bin

Beware of too few events/bin

2) Extends to n dimensions →

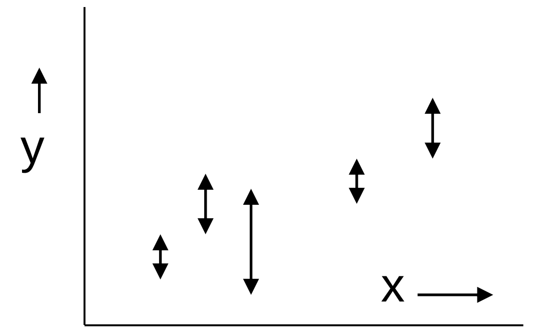
but needs lots of events for n larger than 2 or 3

3) No problem with correlated uncertainties

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

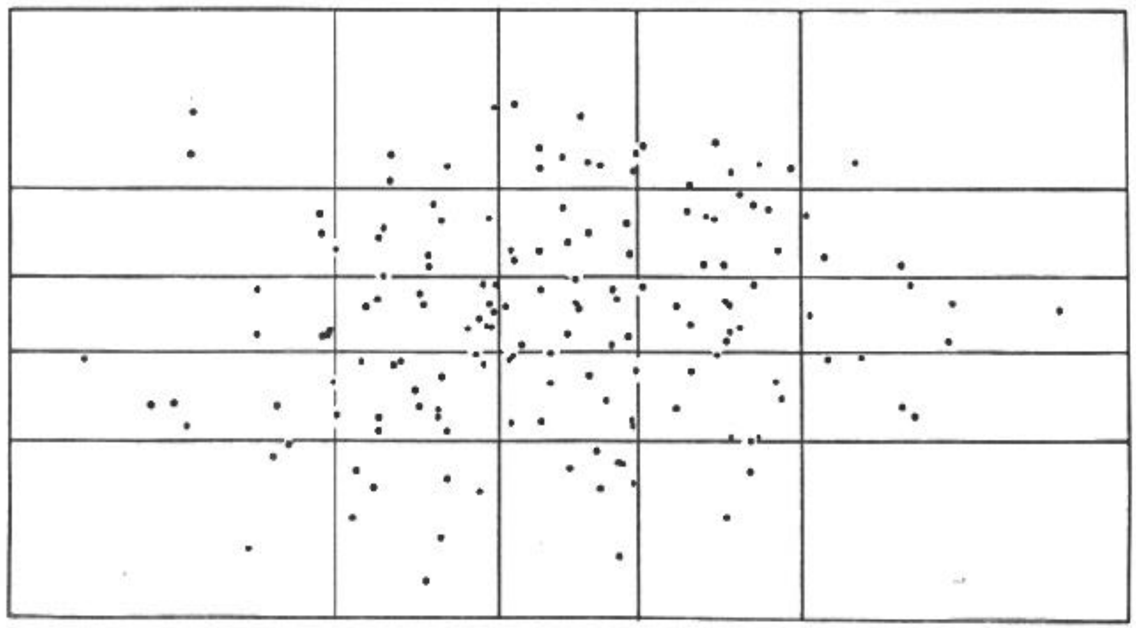
$$\Sigma (y_i - a - bx_i)^2 / \sigma^2 = [y_i^2] - b\,[x_i y_i] - a\,[y_i]$$

5) For theory linear in params, analytic solution



6) Goodness of Fit

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



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

Goodness of Fit: χ^2 test

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

   $\nu = n - p$

   n = no. of data points

   p = no. of FREE parameters
3) Look up probability that, for ν degrees of freedom, $\chi^2 \geq S_{min}$

Uses i) Poisson ~ Gaussian if expected number not too small

ii) For N y_i distributed as Gaussian N(0,1), $\Sigma y_i^2 \sim \chi^2$ with ndf = N

So works ASYMPTOTICALLY. Otherwise use MC for dist of S (or binned $\mathcal{L}$)

Properties of mathematical χ^2 distribution:

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

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

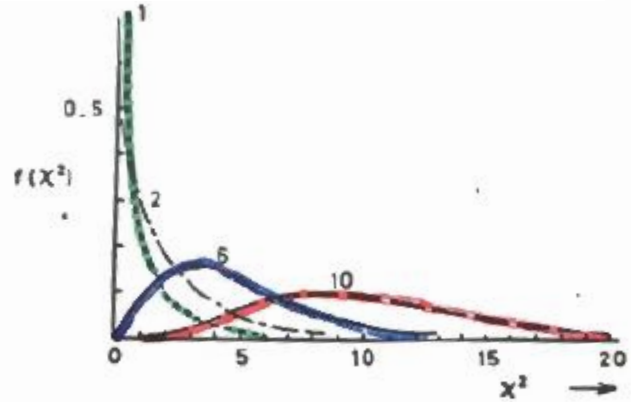
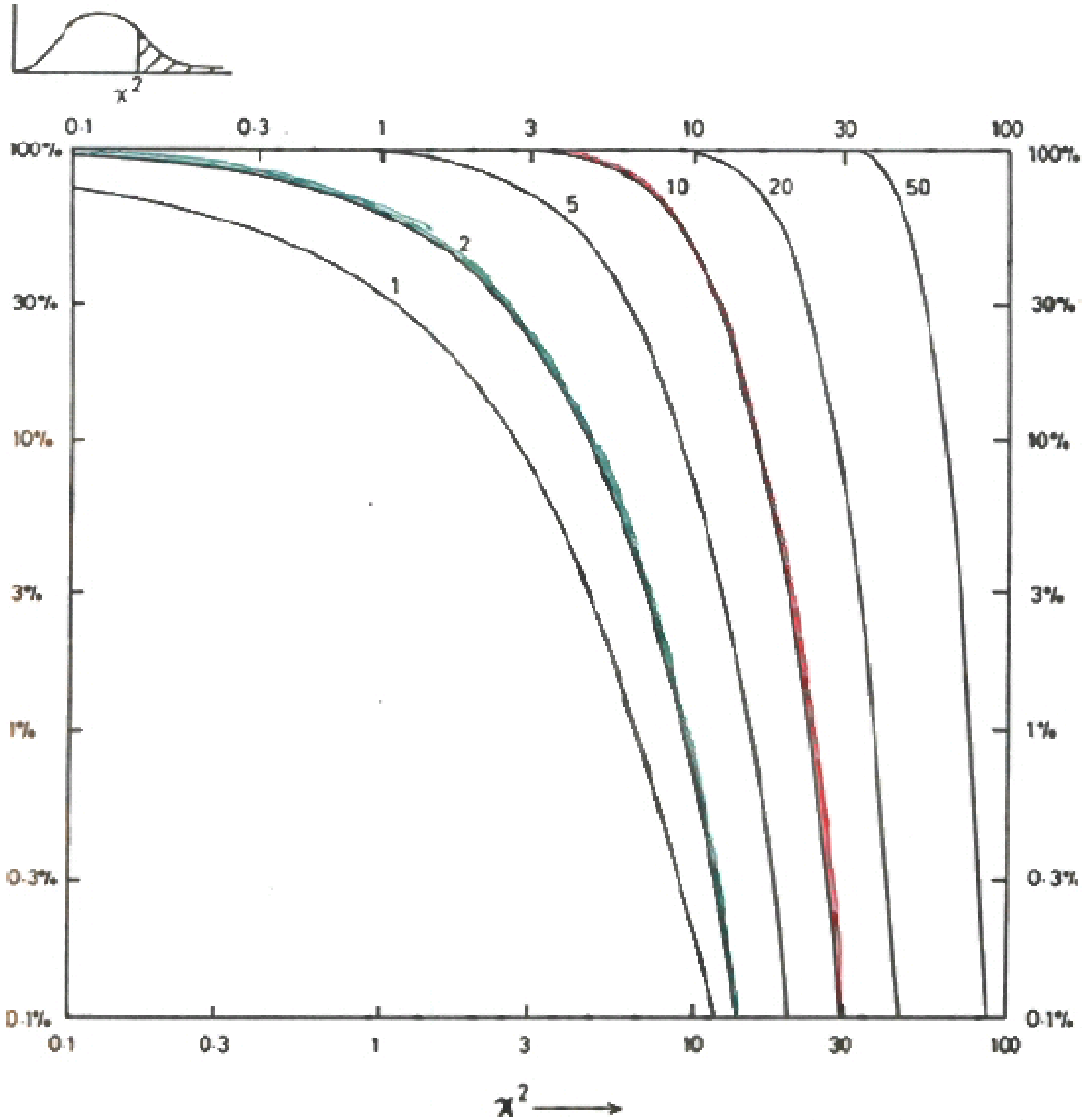


Fig. 2.6

So $S_{min} > \nu + 3\sqrt{2\nu}$ is LARGE

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



Cf: Area in tails of Gaussian

χ^2 with ν degrees of freedom?

ν = data – free parameters ?

Why asymptotic (apart from Poisson → Gaussian) ?

a) Fit flatish histogram with

$y = N \{1 + 10^{-6} \cos(x - x_0)\}$ x_0 = free param

b) Neutrino oscillations: almost degenerate parameters

$y \sim 1 - A \sin^2(1.27\, \Delta m^2\, L/E)$ 2 parameters

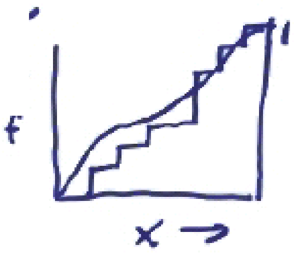
$\xrightarrow{\text{Small } \Delta m^2} 1 - A\, (1.27\, \Delta m^2\, L/E)^2$ 1 parameter

Goodness of Fit

. $\chi 2$ Very general
Needs binning
Not sensitive to sign of deviation

Run Test

Kolmogorov-Smirnov



Aslan and Zech `Energy Test'
Durham IPPP Stats Conf (2002)

Binned $\mathcal{L}$ikelihood (= Baker-Cousins}

etc

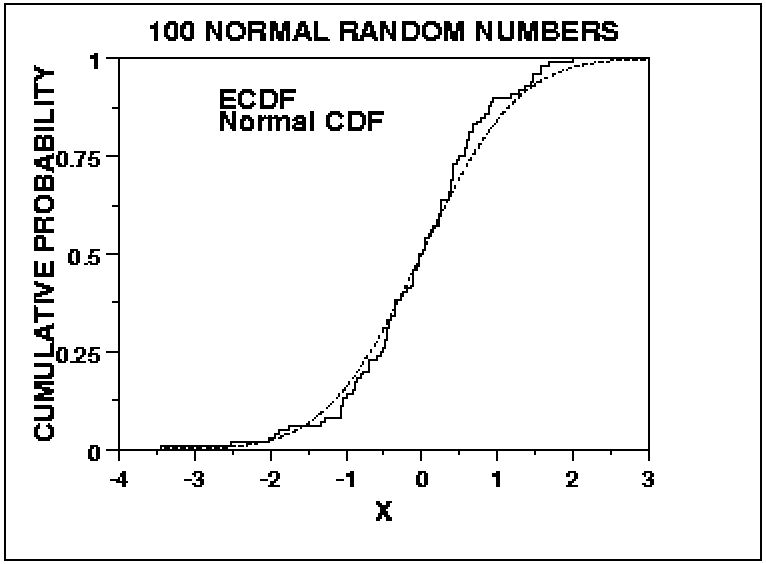
Goodness of Fit: Kolmogorov-Smirnov

Compares data and model cumulative plots

(or 2 sets of data)

Uses largest discrepancy between dists.

Model can be analytic or MC sample



Uses individual data points

Not so sensitive to deviations in tails

(so variants of K-S exist)

Not readily extendible to more dimensions

Distribution-free conversion to p; depends on n

(but not when free parameters involved – needs MC)

Goodness of fit: 'Energy' test

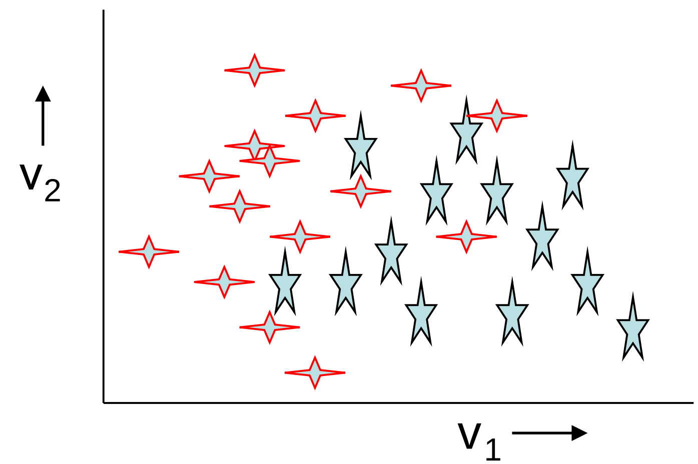
Assign +ve charge to data ; -ve charge to M.C.

Calculate 'electrostatic energy E' of charges

If distributions agree, E ~ 0

If distributions don't overlap, E is positive

Assess significance of magnitude of E by MC



N.B.

1) Works in many dimensions

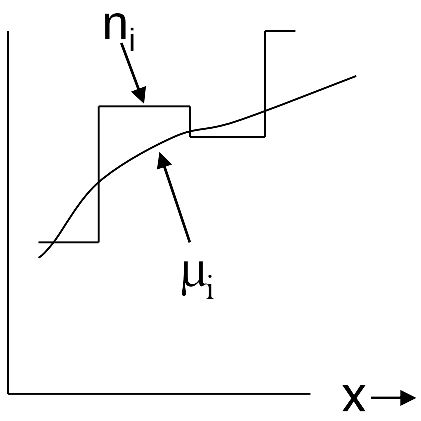
2) Needs metric for each variable (make variances similar?)

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

   Performance insensitive to choice of small ε

See Aslan and Zech's paper at:
http://www.ippp.dur.ac.uk/Workshops/02/statistics/program.shtml

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



For histogram, uses Poisson prob P(n;μ) for n observed events when expect μ.

Construct $\mathcal{L}$-ratio = Product{$P(n_i;\mu_i)/P(n_i;\mu=n_i)$}

$P(n_i;\mu=n_i)$ is best possible μ for that n_i

Need denoms because P(100;100.0) very different from P(1;1.0)

-2*$\mathcal{L}$ ratio ~ χ^2 when μ_i large and $n_i \sim \mu_i$

Better than Neyman or Pearson χ^2 when μ_i small

Baker and Cousins, NIM 221 (1984) 437

Wrong Decisions

Error of First Kind

Reject H0 when true

Should happen x% of tests

Errors of Second Kind

Accept H0 when something else is true

Frequency depends on ....

i) How similar other hypotheses are

e.g. H0 = μ

Alternatives are:	e		π	K	p
ii) Relative frequencies:	10^{-4}	10^{-4}	1	0.1	0.1

Aim for maximum efficiency ⟵ Low error of 1st kind

maximum purity ⟵ Low error of 2nd kind

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

Choose compromise

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

1) Result of experiment

e.g Is spin of resonance = 2?

Get answer WRONG

Where to set cut?

Small cut ⇒ Reject when correct

Large cut ⇒ Never reject anything

Depends on nature of H0 e.g.

Does answer agree with previous expt?

Is expt consistent with special relativity?

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

Error of 1st kind: Loss of efficiency

Error of 2nd kind: More background

Usually easier to allow for 1st than for 2nd

3) Track finding

Combining: Uncorrelated exptl results

Simple Example of Minimising S

Measurements $a_1 \pm \sigma_1$, $a_2 \pm \sigma_2$, ⋮, $a_i \pm \sigma_i$ } 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
$S \to S_{min} + 1$

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

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

Many params

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

N.B. Better to combine data rather than results

So $\hat{a} = \Sigma w_i a_i / \Sigma w_i$, where $w_i = 1/\sigma_i^2$

Difference between weighted and simple averaging

Isolated island with conservative inhabitants

How many married people ?

Number of married men = 100 ± 5 K

Number of married women = 80 ± 30 K

Total = 180 ± 30 K

Wtd average = 99 ± 5 K

→ Total = 198 ± 10 K

CONTRAST

GENERAL POINT: Adding (uncontroversial) theoretical input can improve precision of answer

Compare "kinematic fitting"

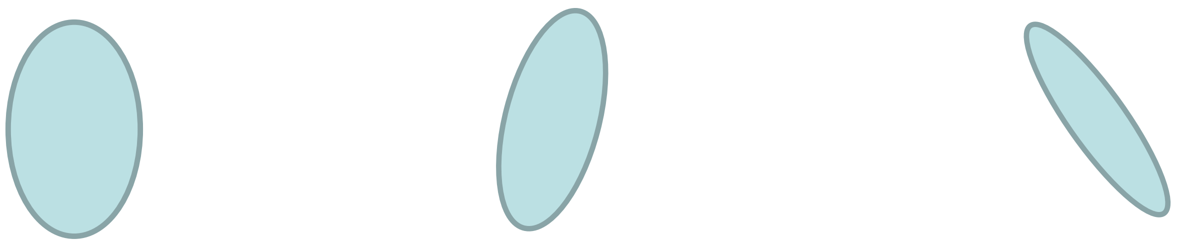
BLUE

Best **L**inear **U**nbiassed **E**stimate

Combine several possibly correlated estimates of same quantity

e.g. v_1, v_2, v_3

Covariance matrix
$$\begin{pmatrix} \sigma_1^2 & cov_{12} & cov_{13} \\ cov_{12} & \sigma_2^2 & cov_{23} \\ cov_{13} & cov_{23} & \sigma_3^2 \end{pmatrix}$$



Uncorrelated   Positive correlation   Negative correlation

$cov_{ij} = \rho_{ij}\, \sigma_i\, \sigma_j$ with $-1 \le \rho \le 1$

Lyons, Gibault + Clifford
NIM A270 (1988) 42

$v_{best} = w_1 v_1 + w_2 v_2 + w_3 v_3$ **L**inear

with $w_1 + w_2 + w_3 = 1$ **U**nbiassed

to give σ_{best} = min (wrt w_1, w_2, w_3) **B**est

For uncorrelated case, $w_i \sim 1/\sigma_i^2$

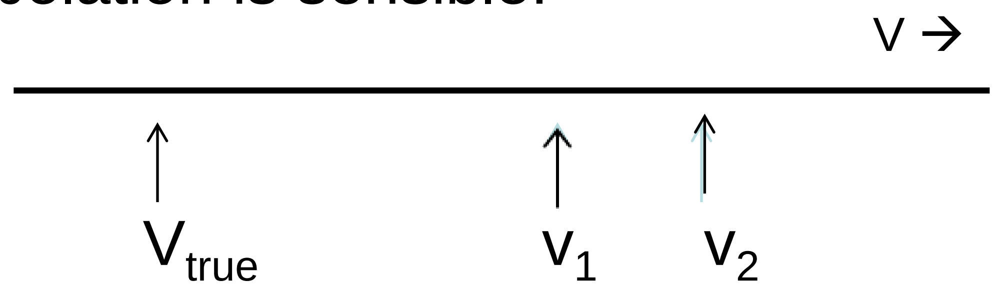
For correlated pair of measurements with $\sigma_1 < \sigma_2$

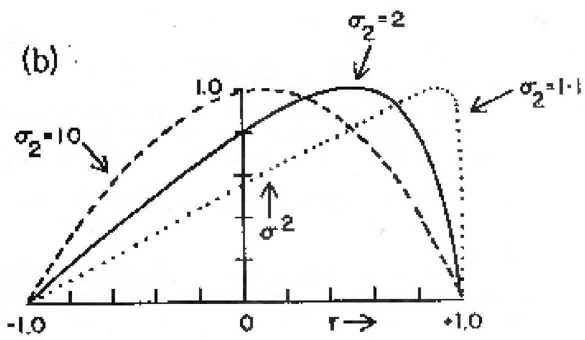
$v_{best} = \alpha v_1 + \beta v_2$ $\beta = 1 - \alpha$

$\beta = 0$ for $\rho = \sigma_1/\sigma_2$

$\beta < 0$ for $\rho > \sigma_1/\sigma_2$ i.e. extrapolation! e.g. $v_{best} = 2v_1 - v_2$

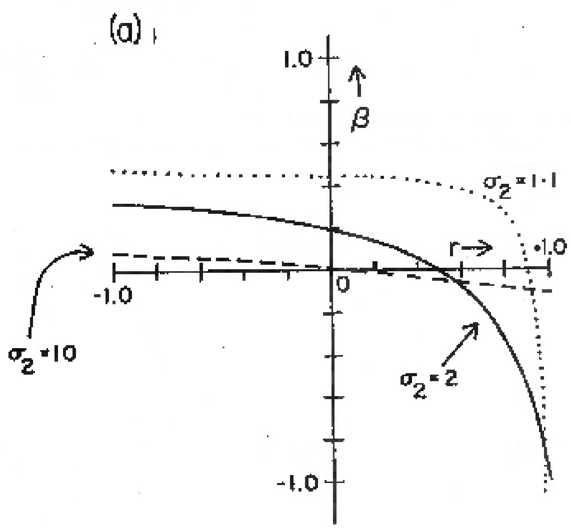
Extrapolation is sensible:





Beware extrapolations because

[b] σ_{best} tends to zero, for ρ = +1 or -1



[a] v_{best} sensitive to ρ and σ_1/σ_2

N.B. For different analyses of ~ same data, ρ ~ 1, so choose 'better' analysis, rather than combining

N.B. σ_{best} depends on σ_1, σ_2 and ρ, but not on $v_1 - v_2$

e.g. Combining 0±3 and x±3 gives x/2 ± 2

BLUE = χ^2

$S(v_{best}) = \Sigma\, (v_i - v_{best})\, E^{-1}_{ij}\, (v_j - v_{best})$, and minimise S wrt v_{best}

S_{min} distributed like χ^2, so measures Goodness of Fit

But BLUE gives weights for each v_i

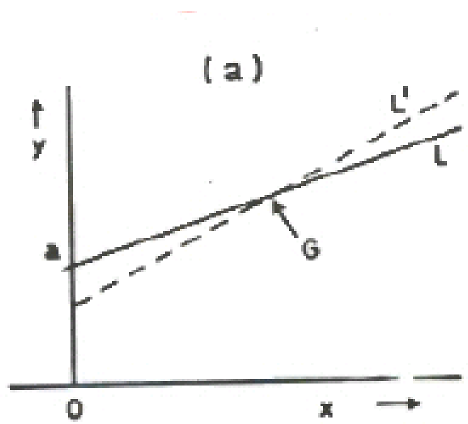
Can be used to see contributions to σ_{best} from each source of uncertainties e.g. statistical and systematics

different systematics

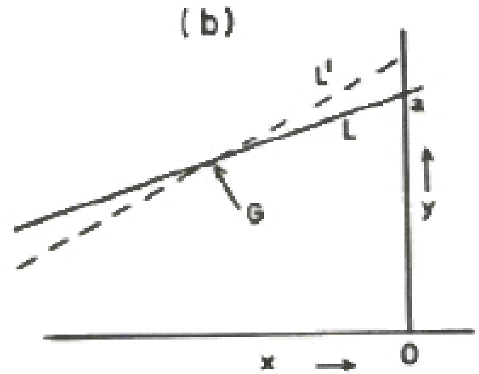
For combining two or more possibly correlated measured quantities {e.g. intercepts and gradients of a straight line), use χ^2 approach.

Alternatively. Valassi has extended BLUE approach

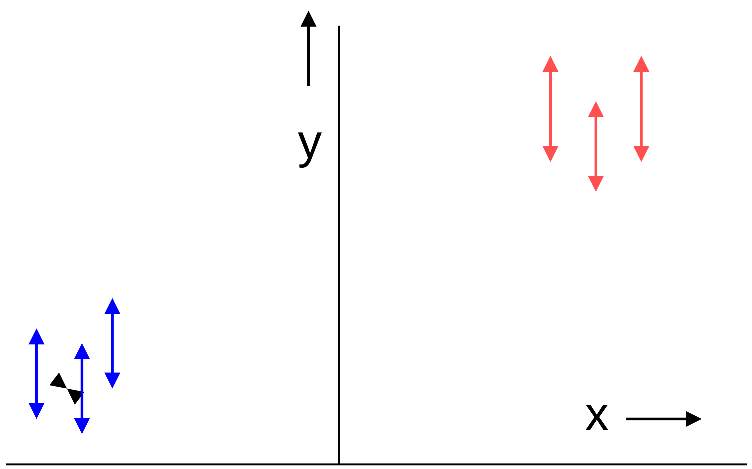
Covariance(a,b) ~ -<x>

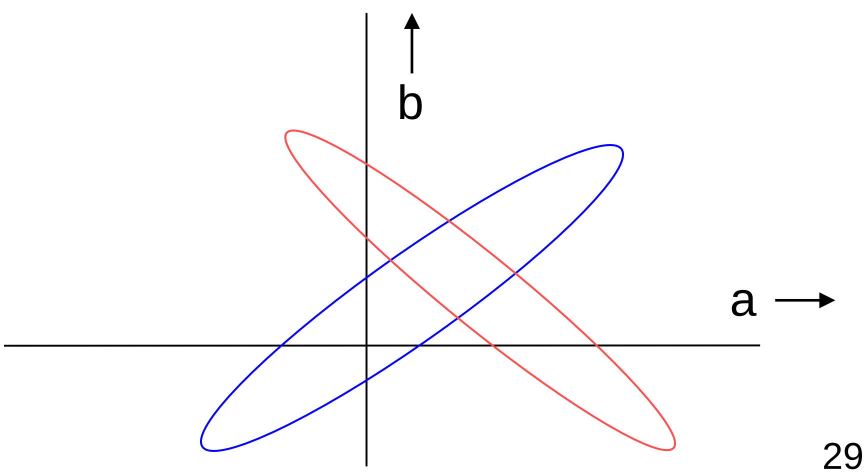


<x> positive



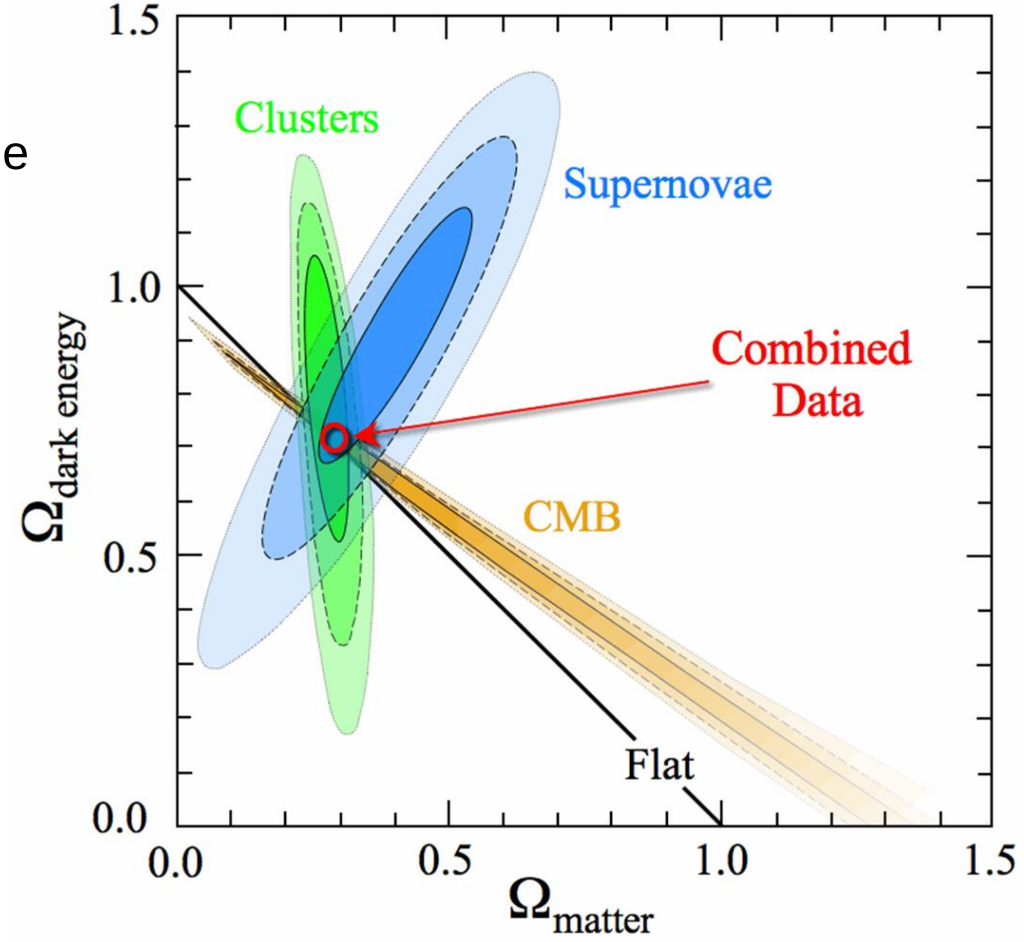
<x> negative





Uncertainty on $\Omega_{\text{dark energy}}$

When combining pairs of variables, the uncertainties on the combined parameters can be **much** smaller than any of the individual uncertainties
e.g. $\Omega_{\text{dark energy}}$



THE PARADOX

Histogram with 100 bins

Fit with 1 parameter

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

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

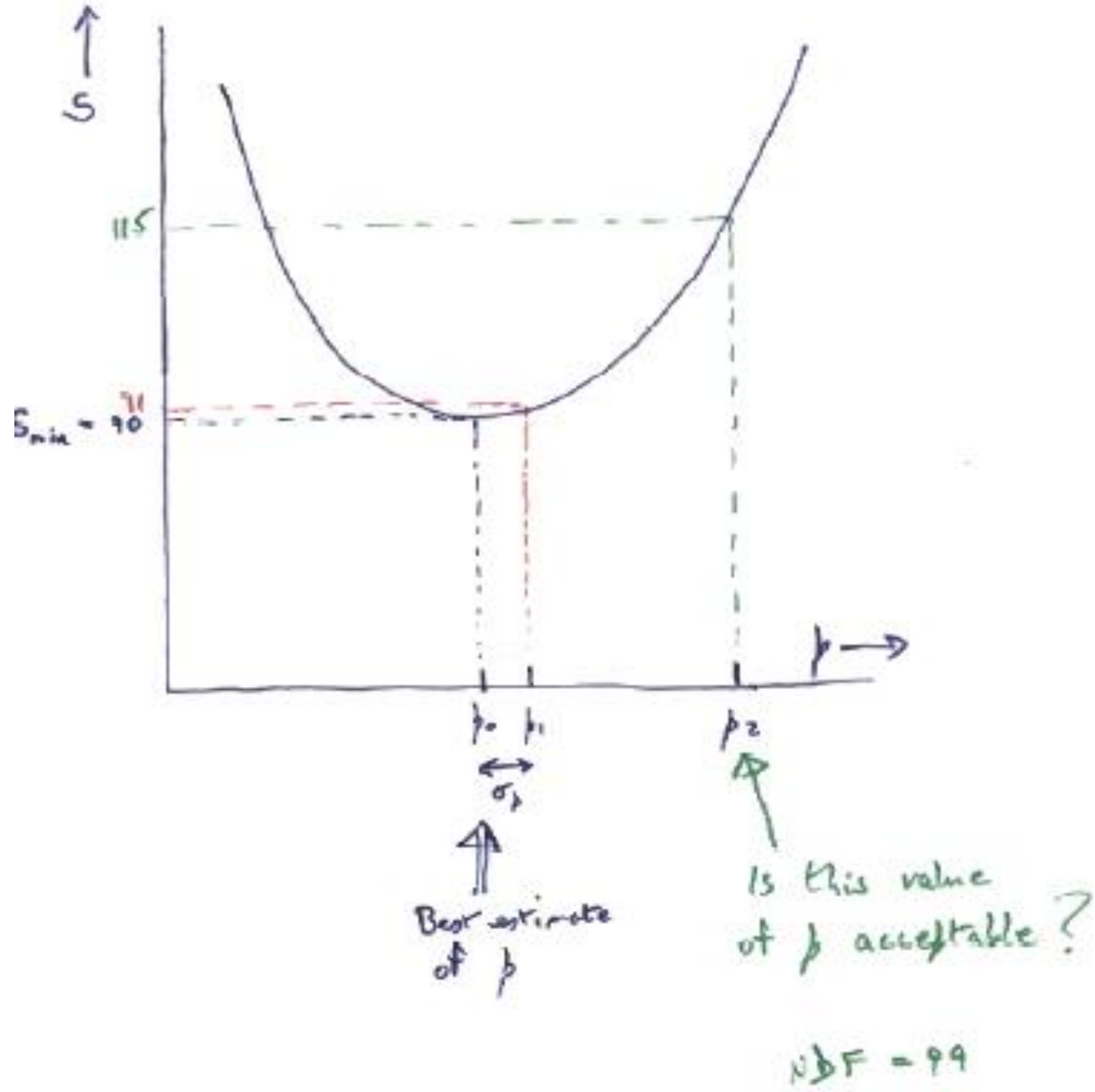
Is p_2 acceptable if $S(p_2) = 115$?

1) YES. Very acceptable χ^2 probability

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

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

So p_2 is 5σ away from best value


S
115
91
S_min = 90
p
p0
p1
p2
σp
Best estimate of p
Is this value of p acceptable?
NDF = 99

Next time:

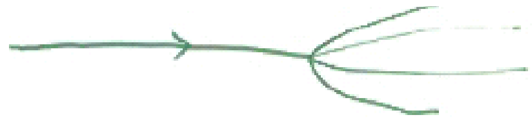
Bayes and Frequentism: the return of an old controversy

The ideologies, with examples
Upper limits
Feldman and Cousins
Summary

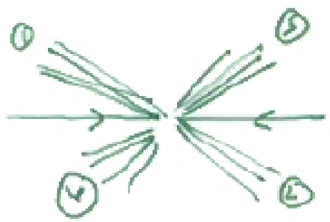
KINEMATIC FITTING

Tests whether observed event is consistent with specified reaction

e.g.



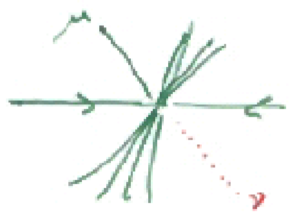
$\bar{p}p \to \bar{p}p\pi^+\pi^-$?



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

$W^+ \hookrightarrow j_1 j_2$, $W^- \to j_3 j_4$

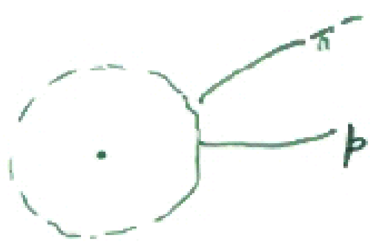
M_W, jet pairings



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

$W^+ \hookrightarrow j_1 j_2$, $W^- \to \mu\nu$

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



$p + \pi^-$ intersect

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

Kinematic Fitting: Why do it?

1) Check whether event consistent with hypothesis [Goodness of Fit]

2) Can calculate missing quantities [Param detn.]

3) Good to have tracks conserving E-P [Param detn.]

4) Reduces uncertainties [Param detn.]

Kinematic Fitting: Why do it?

1) Check whether event consistent with hypothesis [Goodness of Fit]

Use S_{min} and ndf

2) Can calculate missing quantities [Param detn.]

e.g. Can obtain |P| for short/straight track, neutral beam; p_x, p_y, p_z of outgoing ν, n, K^0

3) Good to have tracks conserving E-P [Param detn.]

e.g. identical values for resonance mass from prodn or decay

4) Reduces uncertainties [Param detn.]

Example of "Including theoretical input reduces uncertainties"

How we perform Kinematic Fitting ?

Observed event: 4 outgoing charged tracks
Assumed reaction: $pp \rightarrow pp\pi^+\pi^-$

Measured variables: 4-momenta of each track, v_i^{meas}
(i.e. 3-momenta & assumed mass)
Then test hypothesis:
Observed event = example of assumed reaction

i.e. Can tracks be wiggled “a bit” to do so?
Tested by:

$$S_{min} = \Sigma(v_i^{fitted} - v_i^{meas})^2 / \sigma^2$$

where v_i^{fitted} conserve 4-momenta
(Σ over 4 components of each track)
N.B. Really need to take correlations into account

i.e. Minimisation subject to constraints (involves Lagrange Multipliers)

‘KINEMATIC’ FITTING

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

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

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

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

ALTERNATIVELY:

Sum =183 ± 1.7, while expect 180

Prob{Gaussian 2-tail area beyond 1.73σ} = 8.3%

Toy example of Kinematic Fit

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

+ constraints:

1) Coplanar
2) p_1 at θ_1
3) p_2 at θ_2
4) θ_1 at θ_2 ⇐ 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: $\dfrac{\partial S}{\partial \theta_1} + \lambda \dfrac{\partial C}{\partial \theta_1} = \dfrac{\partial S}{\partial \theta_2} + \lambda \dfrac{\partial C}{\partial \theta_2} = 0$

⇒ 3 eqns for θ_1 θ_2 λ

Eqns simple to solve because

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

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

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

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

i.e. KINEMATIC FIT → REDUCED UNCERTAINTIES