\( \chi^2 \) 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 $\sigma$?

Resume of straight line

Correlated errors

Goodness of fit with $\chi^2$

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
If theory and data OK:

\[ y_{\text{th}} \sim y_{\text{obs}} \rightarrow S \text{ small} \]

Minimise \( S \) \( \rightarrow \) best line

Value of \( S_{\text{min}} \) \( \rightarrow \) how good fit is

\[
S = \sum_{i} \left( \frac{y_{i} - (a + bx_{i}) - y_{\text{obs}}}{\sigma_{i}} \right)^{2}
\]

An error for each pt.
### Which σ should we use?

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

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

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

(Fixed $\sigma_i$)

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

2) Minimise $S$ (w.r.t. $a$, $b$)

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

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

Sim. eqns
For 2 unknowns
($a = b$)

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

$$= \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2}$$

$$a \langle f \rangle = [\langle f \rangle / \langle f \rangle]$$

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

N.B. L.S.B.F. passes through ($\langle x \rangle$, $\langle y \rangle$)
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, \(\text{cov} = 0\)

i.e. gradient and intercept there are uncorrelated

So track params are usually specified at centre of track.
Covariance(a,b) ~ -<x>
Measurements with correlated errors  e.g. systematics?

\[ S = \left( \frac{y - y_{\text{true}}}{\sigma_y} \right)^2 + \left( \frac{x - x_{\text{true}}}{\sigma_x} \right)^2 \]

Introduce correlations by

\[
\begin{align*}
    x &= r \cos \theta - s \sin \theta \\
    y &= + \sin \theta + s \cos \theta
\end{align*}
\]

Not in x-y space

Write \( \sigma_x, \sigma_y \) (4 \( \text{cov}(x,y) = 0 \)) in terms of \( \sigma_x, \sigma_y, \text{cov}(x,y) \)

\[ S = \frac{1}{\sigma_x^2 \sigma_y^2 - \text{cov}(x,y)} \left[ \sigma_x^2 (r - r_{\text{true}})^2 + \sigma_y^2 (s - s_{\text{true}})^2 - 2 \text{cov}(x,y)(r - r_{\text{true}})(s - s_{\text{true}}) \right] \]

Inverse matrix elements

\[ H_{ij} = H_{ij} (r - r_{\text{true}})^2 + H_{ij} (s - s_{\text{true}})^2 + 2 \text{cov}(x,y)(r - r_{\text{true}})(s - s_{\text{true}}) \]

where \( H^{-1} = \left( \begin{array}{cc} \sigma_x^2 & \text{cov} \\ \text{cov} & \sigma_y^2 \end{array} \right) \)

Reduces to standard formula in absence of correlations

In general:

\[ S = \sum_{ij} \Delta_i H_{ij} \Delta_j \]

where \( \Delta_i = (\text{observed} - \text{true})_i \)
Comments on Least Squares method

1) Need to bin
   Beware of too few events/bin

2) Extends to n dimensions
   but needs lots of events for n larger than 2 or 3

3) No problem with correlated uncertainties

4) Can calculate $S_{\text{min}}$ “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

<table>
<thead>
<tr>
<th></th>
<th>Individual events (e.g. in cos $\theta$)</th>
<th>$y_i \pm \sigma_i \ \vee \ x_i$ (e.g. stars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Need to bin?</td>
<td>Yes</td>
<td>No need</td>
</tr>
<tr>
<td>4) $\chi^2$ on line</td>
<td>First histogram</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Moments</td>
<td>Max Like</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Easy?</td>
<td>Yes, if…</td>
<td>Normalisation, maximisation messy</td>
</tr>
<tr>
<td>Efficient?</td>
<td>Not very</td>
<td>Usually best</td>
</tr>
<tr>
<td>Input</td>
<td>Separate events</td>
<td>Separate events</td>
</tr>
<tr>
<td>Goodness of fit</td>
<td>Messy</td>
<td>No (unbinned)</td>
</tr>
<tr>
<td>Constraints</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N dimensions</td>
<td>Easy if ….</td>
<td>Norm, max messier</td>
</tr>
<tr>
<td>Weighted events</td>
<td>Easy</td>
<td>Errors difficult</td>
</tr>
<tr>
<td>Bgd subtraction</td>
<td>Easy</td>
<td>Troublesome</td>
</tr>
<tr>
<td>Inverse covariance matrix</td>
<td>Observed spread, or analytic</td>
<td>$\left{- \frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j}\right}$</td>
</tr>
<tr>
<td>Main feature</td>
<td>Easy</td>
<td>Best</td>
</tr>
</tbody>
</table>
Goodness of Fit: $\chi^2$ test

1) Construct $S$ and minimise wrt free parameters

2) Determine $\nu = \text{no. of degrees of freedom}$
   
   $\nu = n - p$
   
   $n = \text{no. of data points}$
   
   $p = \text{no. of FREE parameters}$

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

Uses
   i) Poisson $\sim$ 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 ASYMPOTICALLY. Otherwise use MC for dist of $S$ (or binned $\mathcal{L}$)
Properties of mathematical $\chi^2$ distribution:

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

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

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

e.g. $S_{\text{min}} = 2200$ for $\nu = 2000$?
Cf: Area in tails of Gaussian
$\chi^2$ with $\nu$ degrees of freedom?

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

Why asymptotic (apart from Poisson $\rightarrow$ Gaussian)?

a) Fit flatish histogram with

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

b) Neutrino oscillations: almost degenerate parameters

$$y \sim 1 - A \sin^2(1.27 \Delta m^2 L/E) \quad 2 \text{ parameters}$$

$$\rightarrow 1 - A (1.27 \Delta m^2 L/E)^2 \quad 1 \text{ parameter}$$

Small $\Delta m^2$
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 Likelihood ( = 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, $E \sim 0$

If distributions don’t overlap, $E$ is positive

Assess significance of magnitude of $E$ by MC

N.B.

1) Works in many dimensions

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

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

Performance insensitive to choice of small $\varepsilon$

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 $P(n;\mu)$ for $n$ observed events when expect $\mu$.

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

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

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

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

Better than Neyman or Pearson $\chi^2$ when $\mu_i$ 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 1\textsuperscript{st} and 2\textsuperscript{nd} kind?

1) Result of experiment
   e.g. Is spin of resonance = 2?
   Get answer WRONG
Where to set cut?
   Small cut $\rightarrow$ Reject when correct
   Large cut $\rightarrow$ Never reject anything
Depends on nature of $H_0$ e.g.
   Does answer agree with previous expt?
   Is expt consistent with special relativity?

2) Class selector e.g. b-quark / galaxy type / $\gamma$-induced cosmic shower
   Error of 1\textsuperscript{st} kind: Loss of efficiency
   Error of 2\textsuperscript{nd} kind: More background
   Usually easier to allow for 1\textsuperscript{st} than for 2\textsuperscript{nd}

3) Track finding
Combining: Uncorrelated expnl results

Simple Example of Minimising $S$

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

N.B. Better to combine data rather than results
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

GENERAL POINT: Adding (uncontroversial) theoretical input can improve precision of answer
Compare “kinematic fitting”
BLUE

Best Linear Unbiassed Estimate

Combine several possibly correlated estimates of same quantity
e.g. $v_1$, $v_2$, $v_3$

Covariance matrix

$$
\begin{bmatrix}
\sigma_1^2 & \text{cov}_{12} & \text{cov}_{13} \\
\text{cov}_{12} & \sigma_2^2 & \text{cov}_{23} \\
\text{cov}_{13} & \text{cov}_{23} & \sigma_3^2
\end{bmatrix}
$$

Uncorrelated  Positive correlation  Negative correlation

$\text{cov}_{ij} = \rho_{ij} \sigma_i \sigma_j$ with $-1 \leq \rho \leq 1$

Lyons, Gibault + Clifford
NIM A270 (1988) 42
\( v_{\text{best}} = w_1 v_1 + w_2 v_2 + w_3 v_3 \) 

Linear

Unbiased

Best

to give \( \sigma_{\text{best}} = \min (\text{wrt } w_1, w_2, w_3) \)

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

For correlated pair of measurements with \( \sigma_1 < \sigma_2 \)

\[ v_{\text{best}} = \alpha v_1 + \beta v_2 \quad \beta = 1 - \alpha \]

\( \beta = 0 \) for \( \rho = \sigma_1/\sigma_2 \)

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

Extrapolation is sensible:

\[ V \rightarrow \]

\[ \uparrow \]

\( V_{\text{true}} \)

\[ \uparrow \quad \uparrow \]

\( v_1 \quad v_2 \)
Beware extrapolations because

[b] $\sigma_{\text{best}}$ tends to zero, for $\rho = +1$ or $-1$

[a] $v_{\text{best}}$ sensitive to $\rho$ and $\sigma_1/\sigma_2$

N.B. For different analyses of ~ same data, $\rho \sim 1$, so choose ‘better’ analysis, rather than combining
N.B. \( \sigma_{\text{best}} \) depends on \( \sigma_1, \sigma_2 \) and \( \rho \), but not on \( v_1 - v_2 \)
e.g. Combining 0±3 and \( x \pm 3 \) gives \( x/2 \pm 2 \)

\[
\text{BLUE} = \chi^2
\]

\[
S(v_{\text{best}}) = \sum (v_i - v_{\text{best}}) E^{-1}_{ij} (v_j - v_{\text{best}}), \text{ and minimise } S \text{ wrt } v_{\text{best}}
\]

\( S_{\text{min}} \) distributed like \( \chi^2 \), so measures Goodness of Fit

But BLUE gives weights for each \( v_i \)

Can be used to see contributions to \( \sigma_{\text{best}} \) 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 \( \chi^2 \) approach.

Alternatively. Valassi has extended BLUE approach
Covariance(a,b) \sim -\langle x \rangle
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
$S_{\text{min}}$: $\chi^2$ with NDF = 99  (Expected $\chi^2 = 99 \pm 14$)

For our data, $S_{\text{min}}(p_0) = 90$
Is $p_2$ acceptable if $S(p_2) = 115$?

1) YES. Very acceptable $\chi^2$ probability

2) NO. $\sigma_p$ from $S(p_0 + \sigma_p) = S_{\text{min}} + 1 = 91$
But $S(p_2) - S(p_0) = 25$
So $p_2$ is 5$\sigma$ away from best value
Is this value of $\beta$ acceptable?

Best estimate of $\beta$

$s_{min} = 90$

$NBF = 99$
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. \( \bar{p}p \rightarrow \bar{p}p \pi^+ \pi^- \) ?

\[ e^+ e^- \rightarrow W^+ W^- \rightarrow j_1, j_2 \rightarrow j_3, j_4 \]

\[ e^+ e^- \rightarrow W^+ W^- \rightarrow j_1, j_2 \rightarrow \mu^+ \mu^- \]

\[ J \rightarrow \bar{p}p^- \text{ from prodn vertex} \]

\[ \bar{p} + \pi^- \text{ interact} \]

\[ \Lambda \rightarrow \bar{p}n^- \text{ from prodn vertex} \]
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 $S_{\text{min}}$ 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 $\nu, n, K^0$

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, \( v_{i}^{\text{meas}} \)
(i.e. 3-momenta & assumed mass)

Then test hypothesis:
Observe event = example of assumed reaction

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

Tested by:

\[
S_{\text{min}} = \sum (v_{i}^{\text{fitted}} - v_{i}^{\text{meas}})^2 / \sigma^2
\]

where \( v_{i}^{\text{fitted}} \) conserve 4-momenta
(\( \sum \) 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: $\theta_1 + \theta_2 + \theta_3 = 180$

\[
\theta_1 \quad \theta_2 \quad \theta_3
\]

Measured 50 60 73±1 Sum = 183

Fitted 49 59 72 180

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

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

ALTERNATIVELY:

Sum = 183 ± 1.7, while expect 180

Prob{Gaussian 2-tail area beyond 1.73$\sigma$} = 8.3%
Toy example of Kinematic Fit

\[ \vec{p}_1 \rightarrow \vec{p}_1 \]

Constraints:
1) Coplanar
2) \( \beta_1 = \theta_1 \)
3) \( \beta_2 = \theta_2 \)
4) \( \theta_1, \theta_2 \) \( \Leftarrow \) Non-relativistic equal mass elastic scatter: \( \theta_1 + \theta_2 = \frac{\pi}{2} \)

Measured: \( \theta_1^m = \sigma \) \( \theta_2^m = \sigma \)
Fitted: \( \theta_1 \) \( \theta_2 \)

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

subject to \( C(\theta_1, \theta_2) = \theta_1 + \theta_2 - \frac{\pi}{2} = 0 \)

Lagrangian: \( \frac{\partial S}{\partial \theta_1} + \lambda \frac{\partial C}{\partial \theta_1} = \frac{\partial S}{\partial \theta_2} + \lambda \frac{\partial C}{\partial \theta_2} = 0 \)

\( \Rightarrow \) 3 eqns for \( \theta_1, \theta_2, \lambda \)
Equis simple to solve because $C(\theta_1, \theta_2)$ linear in $\theta_1$, $\theta_2$

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

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

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

i.e. KINEMATIC FIT $\rightarrow$ REDUCED UNCERTAINTIES