Statistics for HEP

Roger Barlow
Manchester University

Lecture 3: Estimation

About Estimation

Theory

Probability

Calculus

Data

Given these distribution parameters, what can we say about the data?

Given this data, what can we say about the properties or parameters or correctness of the distribution functions?

Statistical

Inference

Data

Theory

What is an estimator?

$$\hat{m}(\{x\}) = \frac{1}{N}\sum_i x_i$$

$$\hat{m}(\{x\}) = \frac{x_{\max} + x_{\min}}{2}$$

$$\hat{V}(\{x\}) = \frac{1}{N}\sum_i (x_i - \hat{m})^2$$

$$\hat{V}(\{x\}) = \frac{1}{N-1}\sum_i (x_i - \hat{m})^2$$

An estimator is a procedure giving a value for a parameter or property of the distribution as a function of the actual data values

What is a good estimator?

One often has to work with less-than-perfect estimators

A perfect estimator is:

- Consistent $\underset{N \to \infty}{Limit}(\hat{a}) = a$
- Unbiassed

$$\langle \hat{a} \rangle = \iiint ... \hat{a}(x_1, x_2, ...) P(x_1; a) P(x_2; a) P(x_3; a) ... dx_1 dx_2 ... = a$$

- Efficient

$$V(\hat{a}) = \left\langle \left(\hat{a} - \langle \hat{a} \rangle\right)^2 \right\rangle$$ minimum

Minimum Variance Bound

$$V(\hat{a}) \geq \frac{-1}{\left\langle \frac{d^2 \ln L}{da^2} \right\rangle}$$

The Likelihood Function

Set of data $\{x_1, x_2, x_3, \ldots x_N\}$

Each x may be multidimensional – never mind

Probability depends on some parameter a

a may be multidimensional – never mind

Total probability (density)

$P(x_1;a)\,P(x_2;a)\,P(x_3;a)\,\ldots P(x_N;a) = L(x_1, x_2, x_3, \ldots x_N\,;a)$

The Likelihood

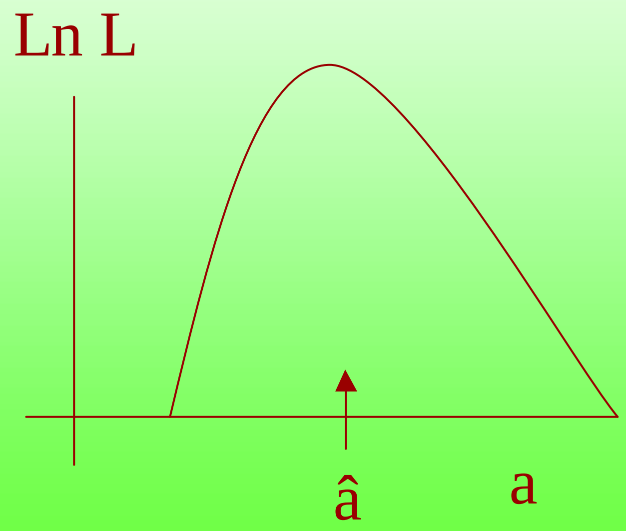
Maximum Likelihood Estimation

Given data $\{x_1, x_2, x_3, \ldots x_N\}$ estimate a by maximising the likelihood $L(x_1, x_2, x_3, \ldots x_N ;a)$

$$\left.\frac{dL}{dA}\right|_{a=\hat{a}} = 0$$

In practice usually maximise ln L as it's easier to calculate and handle; just add the ln $P(x_i)$

ML has lots of nice properties



Properties of ML estimation

- It's consistent

  (no big deal)

- It's biassed for small N

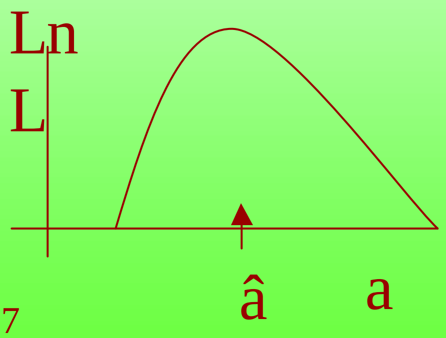
  May need to worry

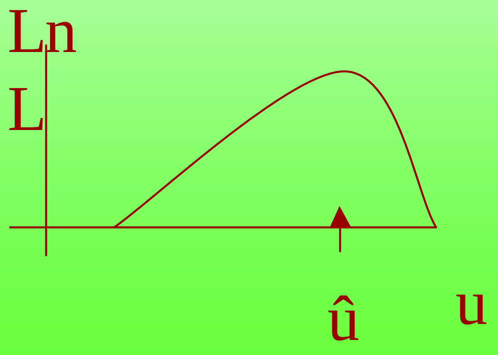
- It is efficient for large N

  Saturates the Minimum Variance Bound

- It is invariant

  If you switch to using u(a), then û=u(â)





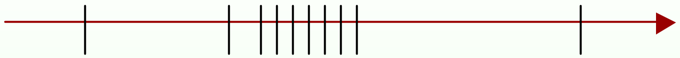
More about ML

- It is not ‘right’. Just sensible.
- It does not give the ‘most likely value of a’. It’s the value of a for which this data is most likely.

- Numerical Methods are often needed
- Maximisation / Minimisation in >1 variable is not easy
- Use MINUIT but remember the minus sign

ML does not give goodness-of-fit

- ML will not complain if your assumed P(x;a) is rubbish
- The value of L tells you nothing

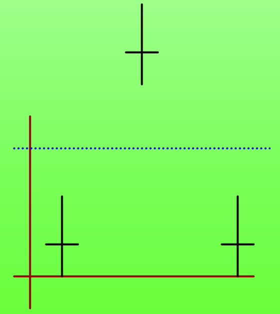


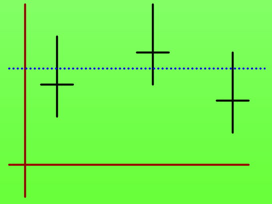
Fit $P(x)=a_1x+a_0$

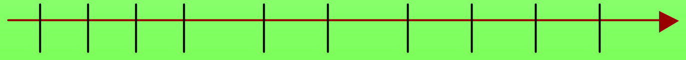
will give $a_1=0$; constant P

$L= a_0^N$

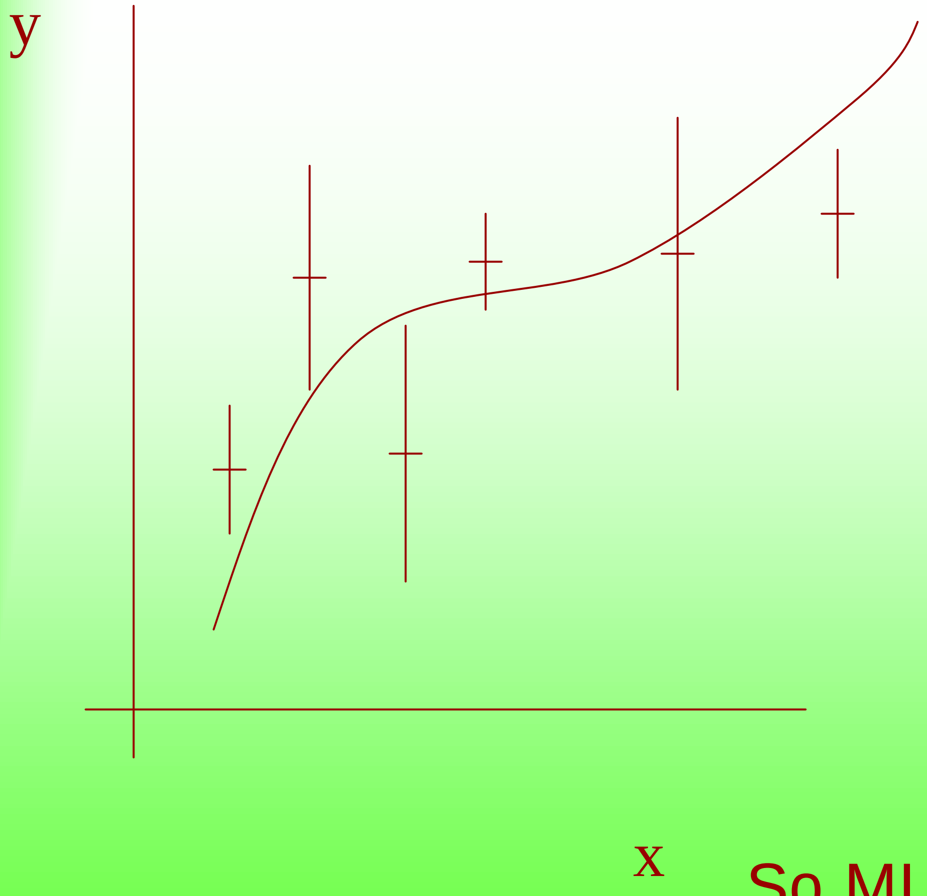
Just like you get from fitting







Least Squares



- Measurements of y at various x with errors σ and prediction f(x;a)
- Probability $\propto e^{-(y-f(x;a))^2/2s^2}$
- Ln L

$$-\frac{1}{2}\sum_i \left(\frac{y_i - f(x_i;a)}{s_i}\right)^2$$

- To maximise ln L, minimise χ^2

So ML ‘proves’ Least Squares. But what ‘proves’ ML? Nothing

Least Squares: The Really nice thing

- Should get $\chi^2 \approx 1$ per data point
- Minimise χ^2 makes it smaller – effect is 1 unit of χ^2 for each variable adjusted. (Dimensionality of MultiD Gaussian decreased by 1.)

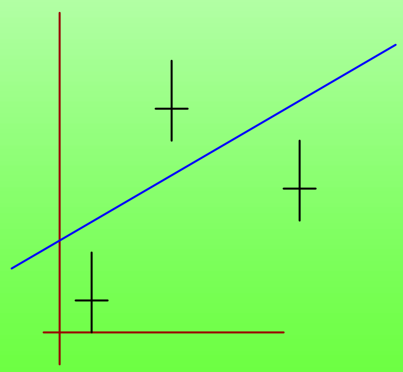
$$N_{\text{degrees Of Freedom}} = N_{\text{data pts}} - N_{\text{parameters}}$$

- Provides 'Goodness of agreement' figure which allows for credibility check

Chi Squared Results

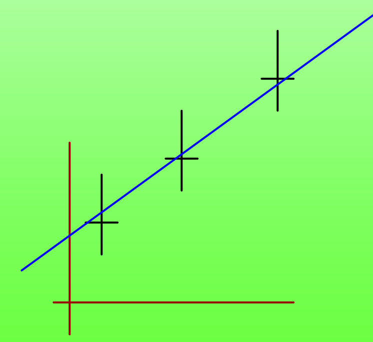
Large χ^2 comes from

1. Bad Measurements
2. Bad Theory
3. Underestimated errors
4. Bad luck



Small χ^2 comes from

1. Overestimated errors
2. Good luck



Fitting Histograms

Often put $\{x_i\}$ into bins

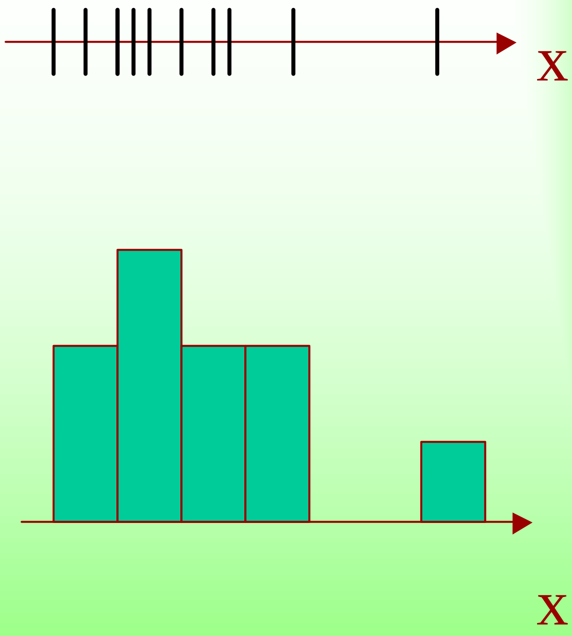
Data is then $\{n_j\}$

n_j given by Poisson,

mean $f(x_j) = P(x_j)\Delta x$

4 Techniques

- Full ML
- Binned ML
- Proper χ^2
- Simple χ^2



What you maximise/minimise

- Full ML $$\ln L = \sum_i \ln P(x_i; a)$$
- Binned ML
$$\ln L = \sum_j \ln Poisson(n_j; f_j) \approx \sum_j n_j \ln f_j - f_j$$
- Proper χ^2 $$\sum_j \frac{(n_j - f_j)^2}{f_j}$$
- Simple χ^2 $$\sum_j \frac{(n_j - f_j)^2}{n_j}$$

Which to use?

- Full ML: Uses all information but may be cumbersome, and does not give any goodness-of-fit. Use if only a handful of events.
- Binned ML: less cumbersome. Lose information if bin size large. Can use χ^2 as goodness-of-fit afterwards
- Proper χ^2 : even less cumbersome and gives goodness-of-fit directly. Should have n_j large so Poisson$\rightarrow$Gaussian
- Simple χ^2 : minimising becomes linear. Must have n_j large

Consumer tests show

- Binned ML and Unbinned ML give similar results unless binsize > feature size
- Both χ^2 methods get biassed and less efficient if bin contents are small due to asymmetry of Poisson
- Simple χ^2 suffers more as sensitive to fluctuations, and dies when bin contents are zero

Orthogonal Polynomials

Fit a cubic: Standard polynomial

$$f(x)=c_0+c_1x+c_2x^2+c_3x^3$$

Least Squares $[\Sigma(y_i-f(x_i))^2]$ gives

$$\begin{pmatrix} 1 & \overline{x} & \overline{x^2} & \overline{x^3} \\ \overline{x} & \overline{x^2} & \overline{x^3} & \overline{x^4} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} & \overline{x^5} \\ \overline{x^3} & \overline{x^4} & \overline{x^5} & \overline{x^6} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \\ \overline{x^2 y} \\ \overline{x^3 y} \end{pmatrix}$$

Invert and solve? Think first!

Define Orthogonal Polynomial

$P_0(x)=1$

$P_1(x)=x + a_{01}P_0(x)$

$P_2(x)=x^2 + a_{12}P_1(x) + a_{02}P_0(x)$

$P_3(x)=x^3 + a_{23}P_2(x) + a_{13}P_1(x) + a_{03}P_0(x)$

Orthogonality: $\Sigma_r P_i(x_r)\, P_j(x_r) = 0$ unless $i=j$

$$a_{ij}=-\left(\Sigma_r\, x_r^{\,j}\, P_i\,(x_r)\right) / \,\Sigma_r\, P_i\,(x_r)^2$$

Use Orthogonal Polynomial

$$f(x)=c'_0P_0(x)+ c'_1P_1(x)+ c'_2P_2(x)+ c'_3P_3(x)$$

Least Squares minimisation gives

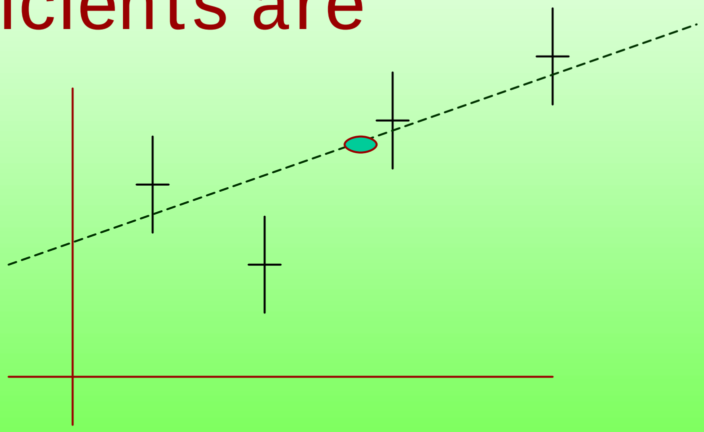
$$c'_i=\Sigma yP_i \;/\; \Sigma P_i^2$$

Special Bonus: These coefficients are UNCORRELATED

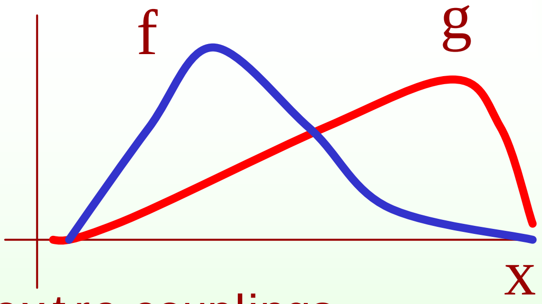
Simple example:

Fit y=mx+c or

$y=m(x - \bar{x})+c'$



Optimal Observables



Function of the form

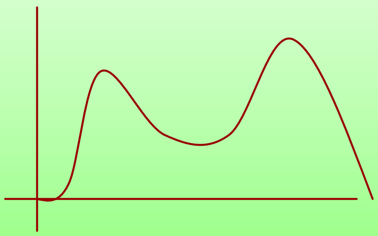
P(x)=f(x)+a g(x)

e.g. signal+background, tau polarisation, extra couplings

A measurement x contains info about a

Depends on f(x)/g(x) ONLY.

Work with O(x)=f(x)/g(x)



Write
$$\overline{O} = \int \frac{f^2}{g} dx + a \int f dx$$

Use
$$\hat{a} = \left(\overline{O} - \int \frac{f^2}{g} dx \right) \Big/ \int f dx$$

Why this is magic

$$\hat{a} = \left(\overline{O} - \int \frac{f^2}{g} dx \right) \Big/ \int f dx$$

It's efficient. Saturates the MVB. As good as ML

x can be multidimensional. O is one variable.

In practice calibrate $\overline{O}$ and â using Monte Carlo

If a is multidimensional there is an O for each

If the form is quadratic then use of the mean OO is not as good as ML. But close.

Extended Maximum Likelihood

- Allow the normalisation of P(x;a) to float
- Predicts numbers of events as well as their distributions

$$N_{pred} = \int P(x;a)dx$$

- Need to modify L

$$\ln L = \sum_i \ln P(x_i;a) - \int P(x;a)dx$$

- Extra term stops normalistion shooting up to infinity

Using EML

- If the shape and size of P can vary independently, get same answer as ML and predicted N equal to actual N
- If not then the estimates are better using EML
- Be careful of the errors in computing ratios and such