Basics 3: Estimation

Roger Barlow
The University of Huddersfield

Terascale Statistics School, DESY, Hamburg

4^{th} July 2023

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

What's happening

You have a dataset $\{x_1, x_2, \dots x_N\}$
and a pdf $P(x, a)$ with unknown parameter(s) a

You want to know:

1. What is the value for a according to the data?
2. What is the error on that value?
3. Does the resulting $P(x, a)$ actually describe the data?

This is called 'estimation' by statisticians and 'fitting' by physicists

Also applies when finding a property rather than a parameter, and then sometimes when one has a parent population rather than a pdf

General considerations

An Estimator is a function of all the x_i which returns some value for a
Write $\hat{a}(x_1, x_2, \ldots x_N)$
There is no 'correct' estimator. You would like an estimator to be

- Consistent: $\hat{a}(x) \to a$ for $N \to \infty$
- Unbiassed: $\langle \hat{a} \rangle = a$
- Efficient: $V(\hat{a}) = \left\langle \hat{a}^2 \right\rangle - \langle \hat{a} \rangle^2$ is small
- Invariant under reparameterisation: $\widehat{f(a)} = f(\hat{a})$
- Convenient

But no estimator is perfect, and these requirements are self-contradictory

Bias: a simple example

Suppose you want to estimate the mean $\mu \equiv \langle x \rangle$ for some pdf, and you choose $\hat{\mu} = \overline{x} = \frac{1}{N}\sum_i x_i$
Then $\langle \hat{\mu} \rangle = \frac{1}{N}\sum_i \langle x_i \rangle = \frac{1}{N}\sum_i \langle x \rangle = \langle x \rangle$. Zero bias.

Suppose you want to estimate the variance $V \equiv \left\langle x^2 \right\rangle - \langle x \rangle^2$ for some pdf, and you choose $\hat{V} = \overline{x^2} - \overline{x}^2 = \frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2$
$\hat{V} = \frac{N-1}{N^2}\sum_i x_i^2 - \frac{1}{N^2}\sum_i \sum_{j \neq i} x_i x_j$
Take expectation values. $\left\langle \hat{V} \right\rangle = \frac{N-1}{N}\left\langle x^2 \right\rangle - \frac{N(N-1)}{N^2}\langle x \rangle^2 = \frac{N-1}{N} V$
The 'obvious' $\hat{V}$ underestimates the true V.

- This is understandable: a fluctuation drags the mean with it, so variations are less
- This can be corrected for (Bessel's correction) by an $N/(N-1)$. Many statistical calculators offer σ_n and σ_{n-1}
- This correction cures the bias for V. Actually σ is still biassed. But V is more useful.
- Biasses are typically small and correctable

Efficiency is limited

Fun algebra with the likelihood function

The Minimum Variance Bound (also called the Cramer-Rao bound)

If $\hat{a}$ is unbiassed (equivalent form exists if it isn't)

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

Start with	Unitarity	No bias
	$\int L(x;a)\,dx = 1$	$\int \hat{a}(x)L(x;a)\,dx = a$
Differentiate	$\int \frac{\partial L}{\partial a}\,dx = 0$	$\int \hat{a}(x)\frac{\partial L}{\partial a}\,dx = 1$
Chain rule	$\int L\frac{\partial \ln L}{\partial a}\,dx = 0^*$	$\int \hat{a}(x)L\frac{\partial \ln L}{\partial a}\,dx = 1$

Multiply column 1 by a and subtract from column 2: $\int (\hat{a}-a)\frac{\partial \ln L}{\partial a} L\,dx = 1$

Invoke Schwarz' lemma $\left(\int uv\,dx\right)^2 \leq \int u^2\,dx \times \int v^2\,dx$

with $u \equiv (\hat{a}-a)\sqrt{L},\ v \equiv \ \frac{\partial \ln L}{\partial a}\sqrt{L}$

$\int (\hat{a}-a)^2 L\,dx. \times \int \left(\frac{\partial \ln L}{\partial a}\right)^2 L\,dx \geq 1$

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

Finally, differentiate Eq. *: $\left\langle \left(\frac{\partial \ln L}{\partial a}\right)^2 \right\rangle + \left\langle \frac{\partial^2 \ln L}{\partial a^2} \right\rangle = 0$ (Fisher information)

Maximum likelihood estimation

The ML estimator

To estimate a using data $\{x_1, x_2 \dots x_N\}$, find the value(s) of a for which the total log likelihood $\sum \ln P(x_i; a)$ is maximised

3 types of problem

1. Differentiate, set to zero, solve the equation(s) algebraically
2. Differentiate, set to zero, solve the equation(s) numerically
3. Maximise numerically

Things to note

- There is no deep justification for ML estimation, except that it works well
- These are not 'the most likely values' of a. They are the values of a for which the values of x are most likely
- The logarithms make the total a sum, which is easier to handle than a product
- Remember a minus sign if you use a minimiser

Maximum likelihood estimation

- Consistent: Almost always
- Unbiassed; It is biassed. But the bias usually falls like $1/N$
- Efficient: In the large N limit ML saturates the MVB, and you can't do better than that
- Invariant under reparameterisation: clearly.
- Convenient. Usually

Simple Examples

$\{x_i\}$ have been gathered from a Gaussian of unknown μ and σ. What are the ML estimates?

$\ln L = \sum -\frac{1}{2}((x_i - \mu)/\sigma)^2 - N \ln(\sqrt{2\pi}\sigma)$

Differentiating wrt μ and σ and setting to zero gives 2 equations

$\sum_i (x_i - \hat{\mu})/\hat{\sigma}^2 = 0 \qquad \sum (x_i - \hat{\mu})^2/\hat{\sigma}^3 - N/\hat{\sigma} = 0$

which are happily decoupled and give

$\hat{\mu} = \frac{1}{N} \sum_i x_i, \qquad \hat{\sigma}^2 = \frac{1}{N} \sum (x_i - \hat{\mu})^2$ (!)

Suppose x_i have been gathered from $P(x; a) = aS(x) + (1 - a)B(x)$

$\ln L = \sum_i \ln(aS(x_i) + (1 - a)B(x_i))$

Differentiate and set to zero

$\sum \frac{S(x_i) - B(x_i)}{\hat{a}S(x_i) + (1 - \hat{a})B(x_i)} = 0$

Needs numerical solution

Errors from ML

To first order, looking at the difference between the true a_0 and the estimated $\hat{a}$

$0 = \frac{\partial \ln L}{\partial a}\Big|_{a=\hat{a}} = \frac{\partial \ln L}{\partial a}\Big|_{a=a_0} + (\hat{a} - a_0)\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}$

Deviations of $\hat{a}$ from a_0 are due to deviations of $\frac{\partial \ln L}{\partial a}\Big|_{a=a_0}$ from zero, divided by the second derivative

$V(\hat{a}) = V(\frac{\partial \ln L}{\partial a}\Big|_{a=a_0}) / \left(\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}\right)^2 = \left\langle \left(\frac{\partial lnL}{\partial a}\right)^2 \right\rangle\Big|_{a=a_0} / \left(\frac{\partial^2 \ln L}{\partial a^2}\Big|_{a=a_0}\right)^2$

Which is all very well, but we don't know what a_0 is...

Approximate by using the actual value of our $\hat{a}$: $V(\hat{a}) = -\left(\frac{\partial^2 \ln L}{\partial a^2}\right)^{-1}$

Noter that this is the MVB (in this approximation). ML is efficient

So the error is given by the second derivative of the log likelihood

---

How to find the second derivative (one way anyway)

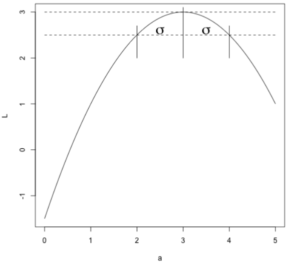
$\ln L(a) = \ln L(\hat{a}) + \frac{1}{2}(a - \hat{a})^2 \frac{\partial^2 \ln L}{\partial a^2} ....$ (first derivative is zero)

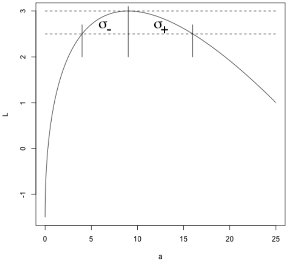
$\quad = \ln L(\hat{a}) - \frac{1}{2}\left(\frac{a - \hat{a}}{\sigma_a}\right)^2$

At $a = \hat{a} \pm \sigma_a$, $\ln L = \ln L(\hat{a}) - \frac{1}{2}$. $\Delta \ln L = -\frac{1}{2}$ gives the error

ML errors

Simple errors
The interval $[\hat{a}-\sigma_a, \hat{a}+\sigma_a]$ from the $\Delta \ln L = -\frac{1}{2}$ points is a 68% central confidence interval





Asymmetric errors (messy!)
If a monotonically reparameterised as $f(a)$, the ML estimate is $\hat{f} = f(\hat{a})$.
$[f(\hat{a}-\sigma_a), f(\hat{a}+\sigma_a)] = [\hat{f}-\sigma_f^-, \hat{f}+\sigma_f^+]$
is the 68% central confidence region.
If $\ln L(a)$ not symmetric parabola, assume this is what is happening and quote separate σ^+, σ^-.

ML errors

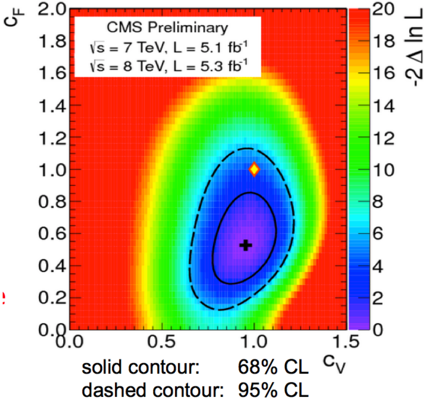
More than one parameter

For 2 (or more) unknown parameters use same technique to map out 68% (or whatever) confidence egions

Only difference is that $\Delta \ln L$ is different.

Given by cumulative probability for χ^2 distribution with 2 (or whatever) degrees of freedom

(Details on χ^2 coming up)



Fitting data points

Suppose your data is a set of x_i, y_i pairs with predictions $y_i = f_i = f(x_i; a)$
x_i known precisely, y_i measured with Gaussian errors σ_i

- Usually one quantity can be precisely specified
- The σ_i may all be the same. If so, the algebra is easier
- The likelihood is the product of Gaussians $\frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}((y_i - f(x_i,a))/\sigma_i)^2}$

$\ln L = -\frac{1}{2}\sum_i (\frac{y_i - f_i}{\sigma_i})^2$+boring constants
Introduce $\chi^2 = \sum_i (\frac{y_i - f(x_i,a)}{\sigma_i})^2$
Maximum Likelihood → minimum χ^2. ('Method of Least Squares')
If f is linear in a (e.g. $f(x) = a_1 + a_2 x + a_3\sqrt{x}$) then this gives a set of equations soluble in one step. If more complicated, need to iterate.

Very simple example: the straight line fit

$$f(x) = a_1 + a_2 x$$

Simple case: all σ_i the same

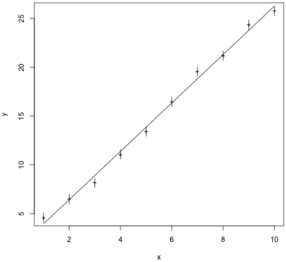
$\chi^2 = \sum \left(\frac{y_i - a_1 - a_2 x_i}{\sigma}\right)^2$

Differentiate and set to zero.

2 Equations

$\sum y_i - a_1 - a_2 x_i = 0$

$\sum x_i(y_i - a_1 - a_2 x_i) = 0$



Simple to unscramble by hand

First is $a_1 = \overline{y} - a_2\overline{x}$

Substitute in 2nd and get $a_2 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$

In more general cases, write these as matrices

Linear Regression

Such straight line fits are linked to the statistical modelling technique of 'linear regression' . The formulæ are the same.

But there are subtle differences

Goodness of fit

Does the model $f(x;a)$ provide a good description of the y_i?

Naïvely each term in χ^2 sum ≈ 1

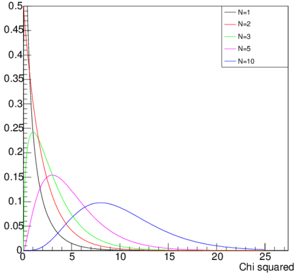
More precisely:

$p(\chi^2, N) = \frac{1}{2^{N/2}\Gamma(N/2)}\chi^{N/2-1}e^{-\chi^2/2}$

Distribution as N dimensional Gaussian, integrated over hypersphere

Quantify by p-value: probability that, if the model is true, χ^2 would be this large, or larger

(p-values apply for any test statistic. Ties up with hypothesis testing. α and p are the same but not the same.)



Each fitted parameter reduces the effective number by 1. (A linear constraint reduces the dimensionality of the hyperspace by 1).

Degrees of freedom $N_D = N - N_f$

Goodness of fit

Reasons for large χ^2:

- Bad theory
- Bad data
- Errors underestimated
- Unsuspected negative correlation between data points (unlikely)
- Bad luck

Reasons for large χ^2:

- Errors overestimated
- Unsuspected positive correlation between data points (more likely)
- Good luck

Although $-\frac{1}{2}\chi^2$ is a log likelihood, $-2\ln L$ is not a χ^2. It tells you nothing about goodness of fit.
(Wilks' theorem says it does for differences in similar models. Useful for comparisons but not absolute.)

4 ways of fitting data

- Full ML. Write down the likelihood and maximise $\sum_j \ln P(x_j, a)$ where j runs over all events. Slow for large data samples, and no goodness of fit.
- Binned ML. Put it in a histogram and maximise the log of the Poisson probabilities $\sum_i n_i \ln f_i - f_i$ where i runs over all bins $f_i = NP(x_i)w$: don't forget the bin width w. Quicker - but lose info from any structure smaller than bin size
- Put it in a histogram and minimise $\chi^2 = \sum_i (n_i - f_i)^2/f_i$ (Pearson's χ^2). This assumes the Poisson distributions are approximated by Gaussians so do not use if bin contents small . But you do get a goodness of fit.
- Put it in a histogram and minimise $\chi^2 = \sum_i (n_i - f_i)^2/n_i$ (Neyman's χ^2). This makes the algebra and fitting a lot easier. But introduces bias as downward fluctuations get more weight. And disaster if any $n_i = 0$

So there are many ways and they are not all equivalent: choose carefully!