Statistics for Particle Physics
Lecture 2: Estimation and Errors

Roger Barlow
The University of Huddersfield
`Roger.Barlow@cern.ch`

LHC Physics school, NCP Islamabad

24th August 2023

Lecture 2: Estimation and Errors

Estimation
statistician-speak for 'Measurement'

The general problem

You know the probability (density) function $P(x; a)$
Take data $\{x_i\}$. What is the best value for a?

x_i may be single values, or pairs, or higher-dimensional
a may be a single parameter or several. If more than one, sometimes split
into 'parameters of interest' and 'nuisance parameters'

Very Broad definition

An Estimator $\hat{a}(x_1 \ldots x_N)$ is a function of the data that gives a value for
the parameter a

A good estimator...

There is no 'correct' estimator - but some are better than others

A perfect estimator would be

Consistent. $\hat{a}(x_1 \ldots x_N) \to a$ as $N \to \infty$
Unbiassed $< \hat{a} >= a$
Efficient $< (\hat{a} - a)^2 >$ is as small as possible
Invariant $\hat{f}(a) = f(\hat{a})$

No estimator is perfect - the goals are incompatible.

The Minimum Variance Bound
also known as the Cramer-Rao bound

Likelihood (again)

$L(a; x_1, x_2, ...x_N) = P(x_1; a)P(x_2; a)...P(x_N; a)$
Probability for the whole data sample, for a particular value of a
Will write $L(a; x_1, x_2, ...x_N)$ as $L(a; x)$ for simplicity

The Minimum Variance Bound (MVB)

For any unbiassed estimator $\hat{a}(x)$, the variance is bounded

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

Maximum Likelihood Estimation

Maximise the likelihood (actually the log likelihood)

$$\text{Maximise } \ln L = \sum_i \ln P(x_i; a) \tag{2}$$

$$\left. \frac{d \ln L}{da} \right|_{\hat{a}} = 0 \tag{3}$$

Is consistent.

Is biassed, but bias falls like $1/N$

Is efficient for large N - saturates the MVB

Is invariant - doesn't matter if you reparameterise a

Particular problem may be solved in three ways depending on complexity

1. Solve Equation 3 algebraically
2. Solve Equation 3 numerically
3. Solve Equation 2 numerically

Least Squares Estimation
Follows from Maximum Likelihood Estimation

Gaussian measurements of y taken at various x values, with measurement error σ, and a prediction $y = f(x; a)$

$$P(y; x, a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-f(x,a))^2/2\sigma^2}$$

$$\ln L = -\sum \frac{(y_i - f(x_i; a))^2}{2\sigma_i^2} + \text{constants}$$

To maximise $\ln L$, minimise $\chi^2 = \sum \frac{(y_i - f(x_i; a))^2}{\sigma_i^2}$

Differentiating gives the Normal Equations: $\sum \frac{(y_i - f(x_i; a))}{\sigma_i^2} f'(x_i; a) = 0$

If $f(x; a)$ is linear in a then these can be solved exactly.
Otherwise use an iterative method.

If the σ_i are all the same, they can be taken out as a common factor

The Straight Line Fit

Function $y = mx + c$

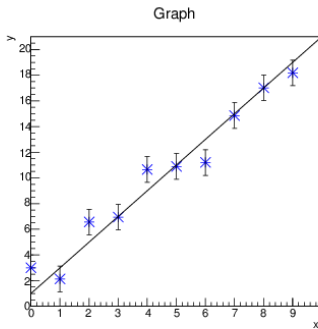Assume all σ_i the same (extension to general case straightforward)

Normal Equations
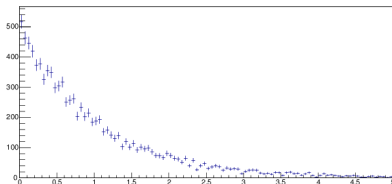
$\sum(y_i - mx_i - c)x_i = 0$

$\sum(y_i - mx_i - c) = 0$

Solution $m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$

$c = \overline{y} - m\overline{x}$

Graph

Fitting Histograms

Fitting a histogram - error given by Poisson statistics so error is $\sqrt{n_i}$

4 methods - increasing accuracy, decreasing speed.

$f_i(x_i; a) = P(x_i; a) \times \text{binwidth}$

1. Minimise $\chi^2 = \sum_i \frac{(n_i - f_i)^2}{n_i}$. Simplest. Breaks if $n_i = 0$

2. Minimise $\chi^2 = \sum_i \frac{(n_i - f_i)^2}{f_i}$ ('Pearson χ^2'. Only for histograms!)

3. Maximise $\ln L = \sum \ln(e^{-f_i} f_i^{n_i}/n_i!) \sim \sum n_i \ln f_i - f_i$. "Binned ML"

4. Ignore bins and maximise likelihood. Sum runs over N_{events} not N_{bins}. Have to use for sparse data.

Errors on estimates
1: Combination of Errors

Well known formula for error on $f(x, y)$ (with x and y independent)

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2$$

If x and y are correlated this becomes

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\rho \left(\frac{\partial f}{\partial y}\right) \left(\frac{\partial f}{\partial x}\right) \sigma_x \sigma_y$$

Generalises to

$$\mathbf{V_f} = \mathbf{G V_x \tilde{G}}$$

where

$$G_{ij} = \frac{\partial f_i}{\partial x_j}$$

You can use this with $f \equiv \hat{a}$

2: Errors from Likelihood

For large N, $\ln L$ curve is parabola
At the maximum,
$\ln L(a) = \ln L(\hat{a}) + \frac{1}{2}(a - \hat{a})^2 \frac{d^2 \ln L}{da^2}$
$-1 / \left\langle \frac{d^2 \ln L}{da^2} \right\rangle$ gives $V(\hat{a})$
(ML saturates MVB)
Approximate $\left\langle \frac{d^2 \ln L}{da^2} \right\rangle \approx \frac{d^2 \ln L}{da^2} \Big|_{a=\hat{a}}$

$\sigma_{\hat{a}} = \sqrt{-\frac{1}{\frac{d^2 \ln L}{da^2}}}$

When $a - \hat{a} = \pm \sigma_{\hat{a}}$,
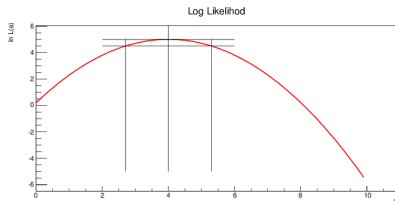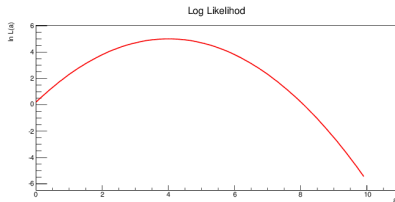$\ln L(a) = \ln L(\hat{a}) - \frac{1}{2}$
Read off errors from $\Delta \ln L = -\frac{1}{2}$
See R.B. arXiv:physics/0403046 (and
Proc Physstat05) for small print

Log Likelihod

Log Likelihod

This gives σ, or 68% errors. Can also take $\Delta \ln L = -2$ for $2\sigma = 95\%$, etc.
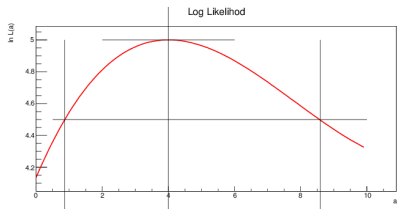If working with χ^2, $L \propto e^{-\frac{1}{2}\chi^2}$ so take $\Delta \chi^2 = 1$

Asymmetric Errors

What if the log likelihood curve is not parabolic? Happens, for instance, in Poisson situations with N not large.:

Say you see 1 event. $\lambda = 1.5$ is more likely to fluctuate down to 1 than $\lambda = 0.5$ fluctuate up to 1.

Log Likelihod

Read off σ_+ and σ_- from the two $\Delta \ln L = -\frac{1}{2}$ crossings

Avoid if possible

Errors in 2 or more dimensions

solid contour: 68% CL
dashed contour: 95% CL

For 2 (or more) dimensions, define regions using contours in $\Delta \ln L$ (or $\Delta \chi^2 \equiv -2\Delta \ln L$)

Levels change:
In 2D, cutting at 1σ square would give $0.68^2 = 47\%$.
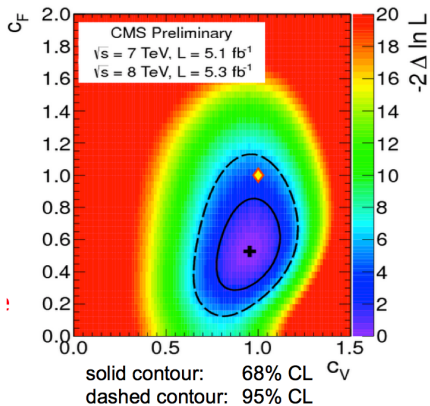A 1σ contour gives 39%.

For 68% use $\Delta \chi^2 = 2.27$
$\Delta \ln L = -1.14$
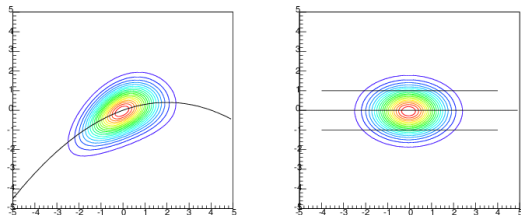For 95% use $\Delta \chi^2 = 5.99$
$\Delta \ln L = -3.00$
(Values from χ^2 distribution)

Profile Likelihood - motivation (not very rigorous)

You have a 2D likelihood plot with axes a_1 and a_2. You are interested in a_1 but not in a_2 ('Nuisance parameter')

Different values of a_2 give different results (central and errors) for a_1

Suppose it is possible to transform to $a_2'(a_1, a_2)$ so L factorises, like the one on the right. $L(a_1.a_2') = L_1(a_1)L_2(a_2')$

Whatever the value of a_2', get same result for a_1

So can present this result for a_1, independent of anything about a_2'.

Path of central a_2' value as fn of a_1, is peak - path is same in both plots

So no need to factorise explicitly: plot $L(a_1, \hat{\hat{a}}_2)$ as fn of a_1 and read off 1D values.

$\hat{\hat{a}}_2(a_1)$ is the value of a_2 which maximises $\ln L$ for this a_1

Instead of profiling, just integrate over a_2.

Can be very helpful alternative, specially with many nuisance parameters

But be aware - this is strictly Bayesian

Frequentists are not allowed to integrate likelihoods wrt the parameter

$\int P(x; a) \, dx$ is fine, but $\int P(x; a) \, da$ is off limits

Reparametrising a_2 (or choosing a different prior) will give different values for a_1

Systematic Errors

There is a lot of bad practice out there. Muddled thinking and following traditional procedures without understanding.

When statistical errors dominated, this didn't matter much. In the days of particle factories and big data samples, it does.

People are ignorant - ignorance leads to fear. They follow familiar rituals they hope will keep them safe.

- What is a Systematic Error?
- How to deal with them
- How to evaluate them
- Checking your analysis
- Conclusions and recommendations

What is a Systematic Error?

Systematic error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique.

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiency with beam position, and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a systematic error.

Bevington

Orear

What is a Systematic Error?

Systematic error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique.

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiency with beam position, and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a systematic error.

Bevington

Orear
These are contradictory

What is a Systematic Error?

Systematic error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique.

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiency with beam position, and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a systematic error.

Bevington

Orear
These are contradictory

Orear is RIGHT

What is a Systematic Error?

Systematic error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique.

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiency with beam position, and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a systematic error.

Bevington

Orear
These are contradictory

Orear is RIGHT

Bevington is WRONG

What is a Systematic Error?

Systematic error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique.

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, variation of counter efficiency with beam position, and energy, dead time, etc. The uncertainty in the estimation of such a systematic effect is called a systematic error.

Bevington

Orear
These are contradictory

Orear is RIGHT

Bevington is WRONG

So are a lot of other books and websites

An error is not a mistake

We teach undergraduates the difference between *measurement errors*, which are part of doing science, and *mistakes*.

If you measure a potential of 12.3 V as 12.4 V, with a voltmeter accurate to 0.1V, that is fine. Even if you measure 12.5 V

If you measure it as 124 V, that is a mistake.

Bevington is describing *Systematic mistakes*

Orear is describing *Systematic uncertainties* - which are 'errors' in the way we use the term.

Avoid using 'systematic error' and always use 'uncertainty' or 'mistake'? Probably impossible. But should always know which you mean

Examples

Track momenta from $p_i = 0.3B\rho_i$ have statistical errors from ρ and systematic errors from B

Calorimeter energies from $E_i = \alpha D_i + \beta$ have statistical errors from light signal D_i and systematic errors from calibration α, β

Branching ratios from $Br = \frac{N_D - B}{\eta N_T}$ have statistical error from N_D and systematics from efficiency η, background B, total N_T

Bayesian or Frequentist?

Can be either

Frequentist: Errors determined by an *ancillary experiment* (real or simulated)

E.g. magnetic field measurements, calorimeter calibration in a testbeam, efficiency from Monte Carlo simulation

Sometimes the ancillary experiment is also the main experiment - e.g. background from sidebands.

Bayesian: theorist thinks the calculation is good to 5% (or whatever). Experimentalist affirms calibration will not have shifted during the run by more than 2% (or whatever)

How to handle them: Correlation

Actually quite straightforward. Systematic uncertainties obey the same rules as statistical uncertainties
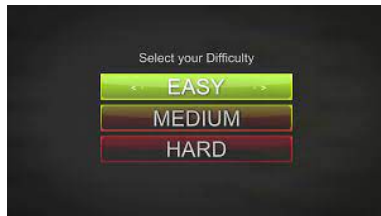
> We write $x = 12.2 \pm 0.3 \pm 0.4$ but we could write $x = 12.2 \pm 0.5$.
> For single measurement extra information is small.

For multiple measurements e.g. $x_a = 12.2 \pm 0.3, x_b = 17.1 \pm 0.4, all \pm 0.5$ extra information important, as results correlated.

Example: cross sections with common luminosity error, branching ratios with common efficiency ...

Handling Systematic Errors in your analysis

3 types

1) Uncertainty in an explicit continuous parameter. E.g. uncertainty in efficiency, background and luminosity in branching ratio or cross section
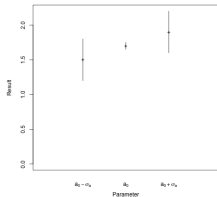
This is trivial. Standard combination-of-errors formula and algebra, just like undergraduate labs

Handling Systematic Errors (2)

Uncertainty in an implicit continuous parameter

Example: MC tuning parameters (σ_{p_T}, polarisation......)

Not amenable to algebra

Method: vary one parameter at a time (OPAT) by $\pm\sigma$ and look at what happens to your analysis result

Note 1: Hopefully effect is equal but opposite - if not then can introduce asymmetric error, but avoid if you can (see earlier).

Note 2. Your analysis results will have errors due to e.g. MC statistics. Some people add these (in quadrature). This is wrong. Technically correct thing to do is subtract them in quadrature, but this is not advised.

Discrete uncertainties, typically in model choice

Situation depends on status of model. Sometimes one preferred, sometimes all equal (more or less)

With 1 preferred model and one other, quote $R_1 \pm |R_1 - R_2|$

With 2 models of equal status, quote $\frac{R_1 + R_2}{2} \pm |\frac{R_1 - R_2}{\sqrt{2}}|$

N models: take $\overline{R} \pm \sqrt{\frac{N}{N-1}(\overline{R^2} - \overline{R}^2)}$ or similar mean value

2 extreme models: take $\frac{R_1 + R_2}{2} \pm \frac{|R_1 - R_2|}{\sqrt{12}}$

These are just ballpark estimates. Do not push them too hard. If the difference is not small, you have a problem - which can be an opportunity to study model differences.

"As we know, there are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know."

Donald H Rumsfeld

Checking the analysis: Errors are not mistakes - but mistakes still happen.

Statistical tools can help find them - though not always give the solution.
Check by repeating analysis with changes which *should* make no difference:

- Data subsets
- Magnet up/down
- Different selection cuts
- Changing histogram bin size and fit ranges
- Changing parametrisation (including order of polynomial)
- Changing fit technique
- Looking for impossibilities
- ...

Example: the BaBar CP violation measurement ".. consistency checks, including separation of the decay by decay mode, tagging category and B_{tag} flavour... We also fit the samples of non-CP decay modes for $\sin 2\beta$ with no statistically significant difference found."

If it passes the test

Tick the box and move on

Do not add the discrepancy to the systematic error

- It's illogical
- It penalises diligence
- Errors get inflated

The more tests the better. You cannot prove the analysis is correct. But the more tests it survives the more likely your colleagues[1] will be to believe the result.

_____

[1] and eventually even you

If it fails the test

Worry!

- Check the test. Very often this turns out to be faulty.
- Check the analysis. Find mistake, enjoy improvement.
- Worry. Consider whether the effect might be real. (E.g. June's results are different from July's. Temperature effect? If so can (i) compensate and (ii) introduce implicit systematic uncertainty)
- Worry harder. Ask colleagues, look at other experiments

  Only as a last resort, add the term to the systematic error. Remember that this could be a hint of something much bigger and nastier

Clearing up a possible confusion

What's the difference between?

> Evaluating implicit systematic errors: vary lots of parameters, see what happens to the result, and include in systematic error

> Checks: vary lots of parameters, see what happens to the result, and don't include in systematic error

(1) Are you expecting to see an effect? If so, it's an evaluation, if not, it's a check

(2) Do you clearly know how much to vary them by? If so, it's an evaluation. If not, it's a check.

Cover cases such as trigger energy cut where the energy calibration is uncertain - may be simpler to simulate the effect by varying the cut.

So finally:

1. Thou shalt never say 'systematic error' when thou meanest 'systematic effect' or 'systematic mistake'.

2. Thou shalt know at all times whether what thou performest is a check for a mistake or an evaluation of an uncertainty.

3. Thou shalt not incorporate successful check results into thy total systematic error and make thereby a shield to hide thy dodgy result.

4. Thou shalt not incorporate failed check results unless thou art truly at thy wits' end.

5. Thou shalt not add uncertainties on uncertainties in quadrature. If they are larger than chickenfeed thou shalt generate more Monte Carlo until they shrink to become so.

6. Thou shalt say what thou doest, and thou shalt be able to justify it out of thine own mouth; not the mouth of thy supervisor, nor thy colleague who did the analysis last time, nor thy local statistics guru

Do these, and thou shalt flourish, and thine analysis likewise.

Backup Slides

Proof of the MVB

Proof.

Unitarity requires $\int P(x; a) \, dx = \int L(a; x) \, dx = 1$

Differentiate wrt a:

$$0 = \int \frac{dL}{da} \, dx = \int L \frac{d \ln L}{da} \, dx = \left\langle \frac{d \ln L}{da} \right\rangle \tag{4}$$

If $\hat{a}$ is unbiased $\langle \hat{a} \rangle = \int \hat{a}(x) P(x; a) \, dx = \int \hat{a}(x) L(a; x) \, dx = a$

Differentiate wrt a: $\quad 1 = \int \hat{a}(x) \frac{dL}{da} \, dx = \int \hat{a} L \frac{d \ln L}{da} \, dx$

Subtract Eq 4 multiplied by a, and get $\int (\hat{a} - a) \frac{d \ln L}{da} L \, dx = 1$

Invoke the Schwarz Inequality $\int u^2 \, dx \int v^2 \, dx \geq \left(\int uv \, dx \right)^2$ with

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

Hence $\int (\hat{a} - a)^2 L \, dx \int \left(\frac{d \ln L}{da} \right)^2 L \, dx \geq 1$

$$\langle (\hat{a} - a)^2 \rangle \geq 1 / \left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle \tag{5}$$

$\square$

Fisher Information

Lemma

Differentiating Equation 4 again gives

$\frac{d}{da} \int L \frac{d \ln L}{da} \, dx = \int \frac{dL}{da} \frac{d \ln L}{da} \, dx + \int L \frac{d^2 \ln A}{da^2} \, dx = \left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle + \left\langle \frac{d^2 \ln L}{da^2} \right\rangle = 0$

Hence $\left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle = - \left\langle \frac{d^2 \ln L}{da^2} \right\rangle$

This is called the Fisher Information. Note how it is intrinsically positive.

Combining Asymmetric Measurements

If you know the likelihood functions, you can do it

Here red and green curves are two measurements of a

To combine them, the log likelihood functions just add (blue)
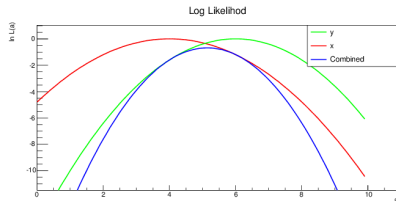

Log Likelihod

But we don't know the full likelihood function: just 3 points (and that it had a maximum at the second)

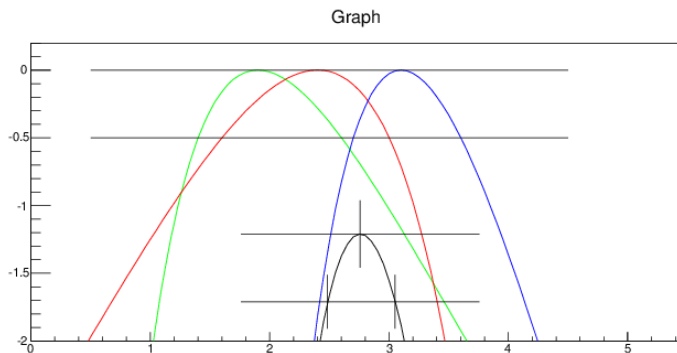Try various models (cubics, constrained quartic...) on likely instances

Two most plausible (for details see RB, arXiv:0406120)

$$\ln L = -\frac{1}{2}\left(\frac{a-\hat{a}}{\sigma+\sigma'(a-\hat{a})}\right)^2 \qquad \ln L = -\frac{1}{2}\frac{(a-\hat{a})^2}{V+V'(a-\hat{a})}$$

Both pretty good. First does better with errors on log a, second does better with Poisson.

Graph

Combining $1.9^{+0.7}_{-0.5}$, $2.4^{+0.6}_{-0.8}$ and $3.1^{+0.5}_{-0.4}$ gives $2.76^{+0.29}_{-0.27}$