Systematic Errors (2)
Working with Systematic Errors

Roger Barlow
Huddersfield University

Terascale Statistics School 2023
DESY, Hamburg

5^{th} July 2023

Why do we quote systematic errors separately?

Results are always given like

In conclusion, we have measured $m = 12.1 \pm 0.3 \pm 0.4$, where the first error is statistical and the second is systematic

Or even '$\pm$ statistical, $\pm$systematic, $\pm$luminosity uncertainty, $\pm$theory uncertainty, $\pm$branching ratio uncertainty'

Why quote them separately?

Why not just 12.1 ± 0.5?

Minor reason - shows whether result is statistics limited
Major reason - to enable combination of this result with others that share a systematic uncertainty

Combination of Errors

What is the error on $f(x, y)$

For undergraduates

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

For graduates

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

If there are several functions and several variables this generalises to

$$\mathbf{V}_f = \tilde{\mathbf{G}} \mathbf{V_x} \mathbf{G} \tag{1}$$

where V_f and V_x are the covariance matrices and $G_{ij} = \frac{\partial f_j}{\partial x_i}$

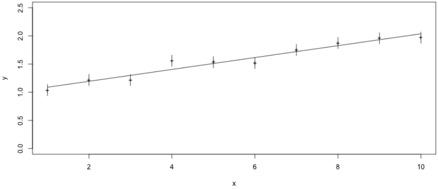
Example - the straight line fit

$y = mx + c$

$$m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sum (x_i - \overline{x}) y_i}{N(\overline{x^2} - \overline{x}^2)}$$

$$c = \overline{y} - m\overline{x} = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} = \frac{\sum (\overline{x^2} - x_i \overline{x}) y_i}{N(\overline{x^2} - \overline{x}^2)}$$

$$\mathbf{V_y} = \sigma^2 \mathbf{I}$$



Equation 1 gives the usual errors, and also the correlation:

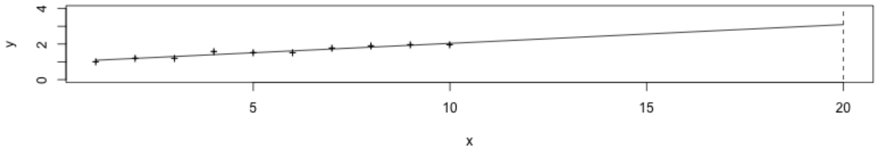
$$V_m = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \qquad V_c = \frac{\sigma^2 \overline{x^2}}{N(\overline{x^2} - \overline{x}^2)} \qquad Cov = -\frac{\overline{x}\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \qquad \rho = -\frac{\overline{x}}{\sqrt{\overline{x^2}}}$$

Note 1: Even though the y_i are independent, m and c are correlated
Note 2: Correlation vanishes if $\overline{x} = 0$. Or write $y = m(x - \overline{x}) + c'$
Note 3: in this example,
$m = 0.105 \pm 0.011, c = 0.983 \pm 0.068, \rho = -0.886$

Example - the straight line fit

Continued

Extrapolation of a straight line - what is y at $x = 20$?



$y = 0.983 + 20 \times 0.105$

Error from $\sqrt{0.068^2 + 20^2 \times 0.011^2} = 0.23$ Wrong

Correct Error from

$\sqrt{0.068^2 + 20^2 \times 0.011^2 - 2 \times 0.886 \times 20 \times 0.068 \times 0.011} = 0.16$

Building a covariance matrix

Matrix element $V_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$

Given correlated x_1 and x_2, model as $x_1 = y_1 + z, x_2 = y_2 + z$, where y_1, y_2, z independent with errors σ_1, σ_2, S. (Example: tracking detector where $y_i \pm \sigma_i$ are the measurements within the detector and $z \pm S$ is the position of the detector.)

$V_{11} = \langle (y_1 + z)(y_1 + z) \rangle - \langle (y_1 + z) \rangle^2 = \sigma_1^2 + S^2$.

V_{22} similar

$V_{12} = V_{21} = \langle (y_1 + z)(y_2 + z) \rangle - \langle (y_1 + z) \rangle \langle (y_2 + z) \rangle = S^2$

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{pmatrix}$$

For more variables, build up larger matrix where off-diagonal elements come from shared features, on-diagonal gives total variance.

Building a correlation matrix

continued

Suppose experiment A measures x_1 and x_2 with shared systematic uncertainty S_A, and experiment B measures x_3 and x_4 with shared S_B

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S_A^2 & S_A^2 & 0 & 0 \\ S_A^2 & \sigma_2^2 + S_A^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 + S_B^2 & S_B^2 \\ 0 & 0 & S_B^2 & \sigma_4^2 + S_B^2 \end{pmatrix}$$

Similar for (more common) shared multiplicative uncertainty - (e.g. efficiency, luminosity, normalisation...)

$x_1 \pm \sigma_1 \pm S_1$ and $x_2 \pm \sigma_2 \pm S_2$ with $S_1 = \xi x_1, S_2 = \xi x_2$

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S_1^2 & S_1 S_2 \\ S_1 S_2 & \sigma_2^2 + S_2^2 \end{pmatrix}$$

PDG, HFLAV and similar groups do this on an industrial scale

Using the matrix

Independent measurements

Maximum Likelihood $\rightarrow$ Least Squares $\rightarrow$ minimise $\chi^2 = \sum_i \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2$

What if the y_i are not independent but correlated with non-diagonal covariance matrix V_y?

Rotate to $\mathbf{y}' = \mathbf{R}\mathbf{y}$ such that $Cov(y'_i y'_j)$ is diagonal

V' diagonal by construction. $\mathbf{V}'^{-1} = \begin{pmatrix} 1/\sigma_1'^2 & 0 & 0 & \dots \\ 0 & 1/\sigma_2'^2 & 0 & \dots \\ 0 & 0 & 1/\sigma_3'^2 & \dots \\ \dots & & & \end{pmatrix}$

and $\mathbf{V}' = \mathbf{R}\mathbf{V}\tilde{\mathbf{R}}$

$\chi^2 = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}})\tilde{\mathbf{R}}[\mathbf{R}\mathbf{V}\tilde{\mathbf{R}}]^{-1}\mathbf{R}(\mathbf{y} - \mathbf{f}) = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f})$

Forget about the primed system and use $\chi^2 = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f})$

The famous Hessian matrix

$$\frac{\partial^2 \ln L}{\partial a_i \partial a_j}$$

$\hat{a_1}$ and $\hat{a_2}$ are functions of the data: maximise $\ln L(a_1, a_2) = \sum_i \ln P(x_i; a_1, a_2)$

To first order about a^{true},

$$\frac{\partial \ln L}{\partial a_1}\Big|_{a=a^{true}} + \frac{\partial^2 \ln L}{\partial a_1^2}(\hat{a_1} - a_1^{true}) + \frac{\partial^2 \ln L}{\partial a_1 \partial a_2}(\hat{a_2} - a_2^{true}) = 0$$

$$\frac{\partial \ln L}{\partial a_2}\Big|_{a=a^{true}} + \frac{\partial^2 \ln L}{\partial a_1 \partial a_2}(\hat{a_1} - a_1^{true}) + \frac{\partial^2 \ln L}{\partial^2 a_2}(\hat{a_2} - a_2^{true}) = 0$$

Same as last lecture on ML errors, but matrix form

Various assumptions (no bias, large N, slow variation so use found values for expectation values...)

$$V_{ij} = -\left\langle \frac{\partial^2 \ln L}{\partial a_j \partial a_k} \right\rangle^{-1}$$

Covariance matrix is just minus the inverse of Hessian matrix, which is (typically) found by minimiser

Averaging
BLUE

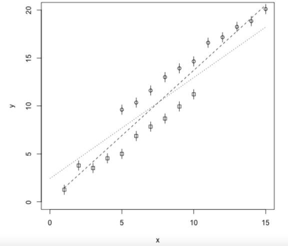
Given several (correlated) results y_i, how do you average them?

Best Linear Unbiased Estimator (L Lyons et al, NIM **A270** 110 (1988))

Minimise $\chi^2 = \sum_{i,j}(y_i - \hat{y}) V^{-1}_{ij} (y_j - \hat{y})$
$\hat{y} \sum_{i,j} V^{-1}_{ij} = \sum_{i,j} V^{-1}_{ij} y_j$
Write as $\hat{y} = \sum_i w_i y_i$ with $w_i = \frac{\sum_j V^{-1}_{ij}}{\sum_{i,j} V^{-1}_{ij}}$

Error on $\hat{y}$ given by $\sqrt{\tilde{\mathbf{w}}\mathbf{V}\mathbf{w}}$
Notice that $\sum_i w_i = 1$ which is intuitive
Notice that some w_i may be negative (if correlations are large) which is counterintuitve
This assumes the elements of $\mathbf{V}$ are known exactly. If not, care needed.

Equivalent alternative for additive systematics



Obvious method: Construct full covariance matrix **V** and minimise χ^2

Alternative: introduce explicit offsets $y'_{ij} = y_{ij} + \xi_j$ for value i of expt j.

ξ_j Gaussian with mean 0, sd S_j, included in χ^2

Fit the ξ_i and the parameter(s) a

Downside: n more parameters to fit

Upside (1) avoids matrix inversion

Upside (2): extracts the factors which can be useful to check behaviour

These two methods are actually (surprisingly!) equivalent

RB. *Combining experiments with systematic errors* **NIM A 987** 164864 (2021)

Also works for multiplicative systematics

And avoids "D'Agostini bias"
G. D'Agostini NIM **A346** 306 (1994)

In combining experiments adjust parameter(s) a to minimise
$\chi^2 = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}}(x;a))\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f}(x;a))$

If **V** includes multiplicative systematic errors (from normalisation) this leads to bias
$S_i = \xi y_i$ so small y_i have increased weight to lower χ^2

Separate fit to systematic factors and applying to the f_i avoids this (at the cost of more complicated solution)

Nuisance Parameters

Another way of thinking about systematic errors.

Suppose you have a joint likelihood function for parameters a_1 and a_2 - perhaps N_S and N_B

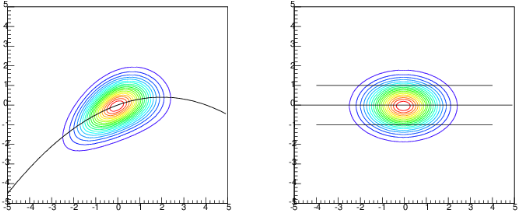
But a_2 is of no interest

Can fix a_2 with some uncertainty (systematic error)

Or can call it a nuisance parameter and get rid of it, by profiling or marginalisation

Nuisance Parameters I

Profile Likelihood - motivation (not very rigorous)



2D likelihood plot with axes a_1 (interesting) and 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

Note how the profile likelihood is a bit broader than a slice at constant a_2

Nuisance Parameters 2

Marginalised likelihoods

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. With a bit of luck, even radical changes in the prior for a_2 will not effect the frequentist result for a_1.
But don't just leave it to luck. Check and make sure.

Conclusions

Systematic errors can readily be handled - with the help of the correlation matrix and other techniques