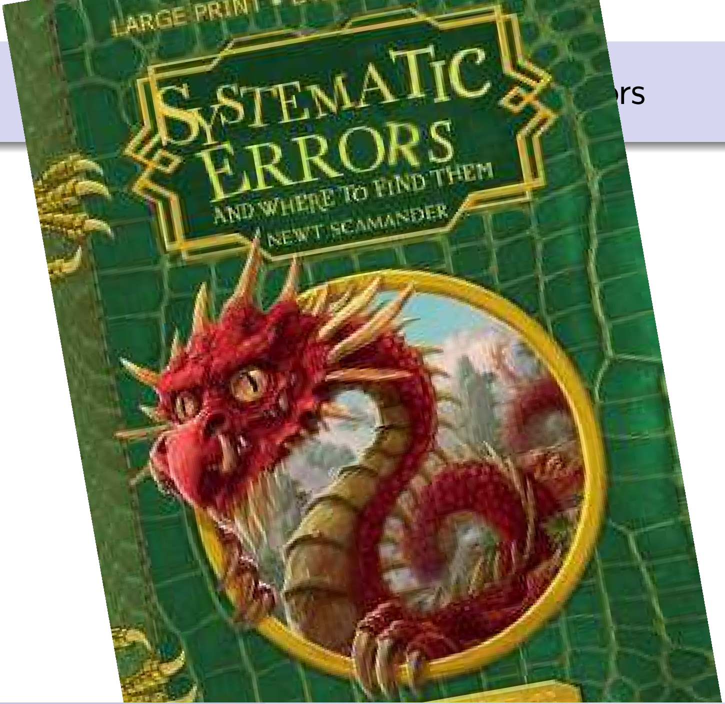
Handling Systematic Errors

Roger Barlow
The University of Huddersfield

STAR collaboration meeting - Juniors' day

1^{st} March 2021

rs

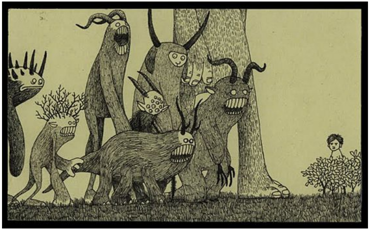

LARGE PRINT
SYSTEMATIC ERRORS
AND WHERE TO FIND THEM
NEWT SCAMANDER

Systematic Errors

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

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

People are scared of systematic errors because they 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.

Bevington

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.

Orear

What is a Systematic Error?

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

Bevington

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.

Orear

These are contradictory

What is a Systematic Error?

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

Bevington

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.

Orear

These are contradictory

Orear is RIGHT

What is a Systematic Error?

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

Bevington

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.

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.

Bevington

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.

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

Nuisance Parameters

is a useful way to think about systematic uncertainties. Parameters which are unknown and hence uncertain, but that you're not interested in.

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)

Some analysis techniques use hybrid of frequentist and Bayesian.

How to handle them: Correlation

Actually straightforward. Systematic uncertainties obey the same rules as statistical uncertainties

We write $x = 12.2 \pm 0.3 \pm 0.4$ "where the first is the statistical and the second is the systematic error", 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 ...

Taking more measurements and averaging does not reduce the error.

Combining results

Correlation expressed through covariance (or variance) matrix

$V_{ij} = < x_i x_j > - < x_i > < x_j > = \delta_{ij}\sigma_i^2 + S^2$

then use $\mathbf{V}^{-1}$ in χ^2: $\sum_i \frac{(x_i - f_i)^2}{\sigma_i^2} \to \sum_{i,j}(x_i - f_i)(x_j - f_j)V_{ij}^{-1}$

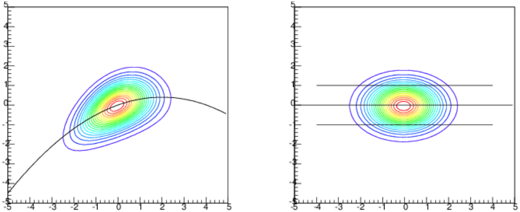
Works for both additive and multiplicative errors

Can write down and invert full matrix, or can introduce parameters to describe the shifts. Methods turn out to be equivalent.

(RB: **NIM A 987** 164184 (2021))

Non-Gaussian errors I

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

Non-Gaussian errors 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

Evaluating Systematic Errors in your analysis

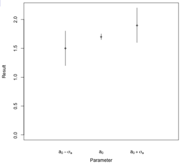
3 types

1) Uncertainty in an explicit continuous parameter:

E.g. uncertainty in efficiency, background and luminosity in branching ratio or cross section

Standard combination of errors formula and algebra, just like undergraduate labs. Have to include correlations but this is all handled by matrices.

Handling Systematic Errors (2)



Uncertainty in an implicit continuous parameter such as: MC tuning numbers (σ_{p_T}, polarisation......)

Not amenable to algebra

Method: vary parameter by $\pm\sigma$ and look at what happens to your analysis result (directly, or through efficiency, background etc.)

Note 1: Hopefully effect is equal but opposite - if not then can introduce asymmetric error, but avoid if you can. Rewrite $^{+0.5}_{-0.3}$ as ± 0.4

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.

Note 3: Or take many Gaussian samples of parameter value and look at distribution of result. Nice, if you have the computing capacity.

Handling Systematic Errors (3)

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 $R_1 \pm |R_1 - R_2|$

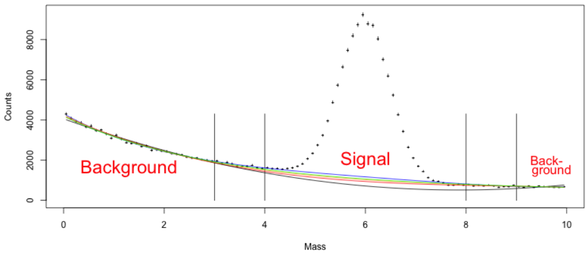
With 2 models of equal status, quote $\frac{R_1+R_2}{2} \pm |\frac{R_1-R_2}{\sqrt{2}}|$

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.

Example: many models



Analysis

Count number of events in signal region.
Subtract background by fitting quadratic in background region (red curve)

Systematics

Try using cubic, quartic, exponential, exponential-quadratic ... functions to fit background

Points to note

- The background function has no theoretical grounding. It just fits the data.
- Sometimes signal and background regions can be in different channels/experiments, but same logic applies.
- For the sake of the example, we suppose all functions give acceptable fits to data in the background region.
- If you know the likelihood function (including the peak) you can use the profile+envelope method [1], but not if you're just counting numbers

[1] P. D. Dauncey, M. Kenzie, N. Wardle and G. J. Davies, Journal of Instrumentation 10 p04015 (2015), arXiv:1408.6865v5

Possible choices

For your central analysis value, you can stick with the quadratic or you can use the average of all functions, depending on what you think of their status.

For the uncertainty you can take the RMS deviation of the values about the chosen central value. Or the mean absolute deviation.

If you take it as the maximum deviation then this is the wrong choice.

Why this is wrong

- It is over-conservative and inflates your errors
- It doesn't really represent a set of values. Mean, mode, median - but max!!
- It penalises diligence: if you consider many functions you are bound to make your errors larger

When is it correct to take the maximum deviation?

Very seldom

- If you're an engineer. Then you quote tolerances. If you have a 99 ± 1 mm peg and a 100 mm hole you want it to fit *every* time.
- If you define 'error' as '68% central confidence region' as opposed to 'rms spread'[2] and take an ultrafrequentist approach to '68% confidence means θ lies in the region at least 68% of the time', i.e. for absolutely all values of ν, rather than profiling or even marginalising. Note that if you take this route you can no longer add in quadrature

We must not be afraid of quoting a result that may be more than 1 sigma from the true value.

[2] For Gaussian distributions it's the same

Why do people do it?

"It's what we always do"

That is not a valid reason for doing anything

"If the result turns out to be outside the quoted error will be bad for my/our reputation"

32 % of our results should be outside their quoted error

"It's conservative"

For errors, *conservative* is another word for *wrong*

Checking the analysis

"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 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[3] will be to believe the result.

[3] and eventually even you

If it fails the test

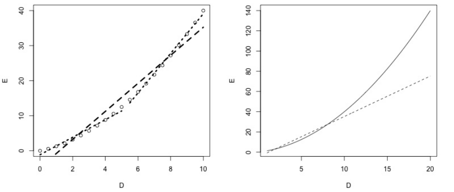
Worry!

- 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

Why failed tests need careful handling

Something bigger and nastier



You have an energy calibration (circles - errors suppressed for clarity).

It is quadratic, but you fit it with a straight line (dashed) $m = 4.00$.

As a check, you fit low values and high values separately (dotted lines) and get different results. (Slopes are 2.50 and 5.65)

You fail to spot the inadequacy of linear calibration, and assign a systematic uncertainty of 1.6 to the slope.

This is a terrible decision. For interpolation it is far too large: the $m = 4$ fit is not that bad. For extrapolation (2nd plot) it is far too small.

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, nor thy mate down the pub.

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