Systematic Errors: Facts and Fictions

Roger Barlow

THE UNIVERSITY of MANCHESTER

I Errors: Mistakes and Uncertainties

– Diversion: Effect of different priors

II Evaluation of Uncertainty

– Diversion: The Errors on Errors Puzzle

III Checking for Mistakes

– Diversion: Defining a ‘Small Difference’

IV Systematics in practice

I: Rival Definitions: Uncertainty and Mistake

'*Systematic effects* is a general category which includes effects such as background, selection bias, scanning efficiency, energy resolution, angle 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

'*Systematic Error:* reproducible inaccuracy introduced by faulty equipment, calibration or technique.' - Bevington

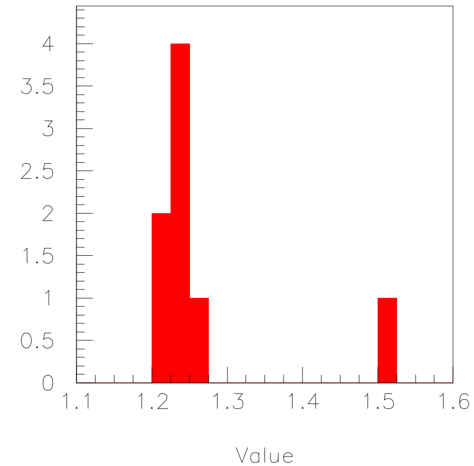
Examples 1:

- Calorimeter energy E from digitisation D: $E = \alpha D$
- Branching ratio B from number of decays seen N : $B = N/\eta N_T$

Examples 2:

- Forgetting to allow for thermal expansion of steel rule
- Rounding down numerical values

Touch base: Random Uncertainties and Mistakes



In some readings (of the same quantity)

1.23, 1.25, 1.24, 1.25, 1.21, 1.52, 1.22, 1.27

you can see some *uncertainty* and a *mistake*.

Statistics provides tools to identify and use an uncertainty

Statistics provides tools to identify a mistake, but not to use it.

Semantics

Physicists generally use *random error* to denote *random uncertainty* not *random mistake.*

For consistency we should use *systematic error* in the same way.

That means Orear's definition, not Bevington's.

And that we must distinguish systematic *Effects* from the *Errors* which are the *Uncertainties* in those effects.

Systematic Mistakes still need to be identified.

Calling them by their proper name makes clear that Statistics doesn't provide tools to tell us what to do with them.

Systematic Errors and Bias

Treated as synonymous by some authors

(e.g. van der Waerden 'The bias, or systematic error, of the estimator...', Kendall & Buckland, Fisher...)

This is not enough. Need to distinguish between:

i) We know the bias, and remove it. End of story.

ii) We don't realise that the bias exists, so we do nothing. This is a mistake.

iii) We know that the bias exists, but not its sign or magnitude.

Examples: expanding steel rules, rounding digits.

Systematic Errors can be Bayesian

Many measurements $\rightarrow$ same result each time Not very frequentist!

May be different ensemble: Component spread, Calibration experiment

May be no escape:

e.g. Luminosity Measurement. Bhabha cross section calculated exactly to α^3.

Calculation always gives same result (with same inaccuracy).

Can guess at this inaccuracy. This is a subjective (Bayesian) probability.

Prior Pitfalls: an illustration

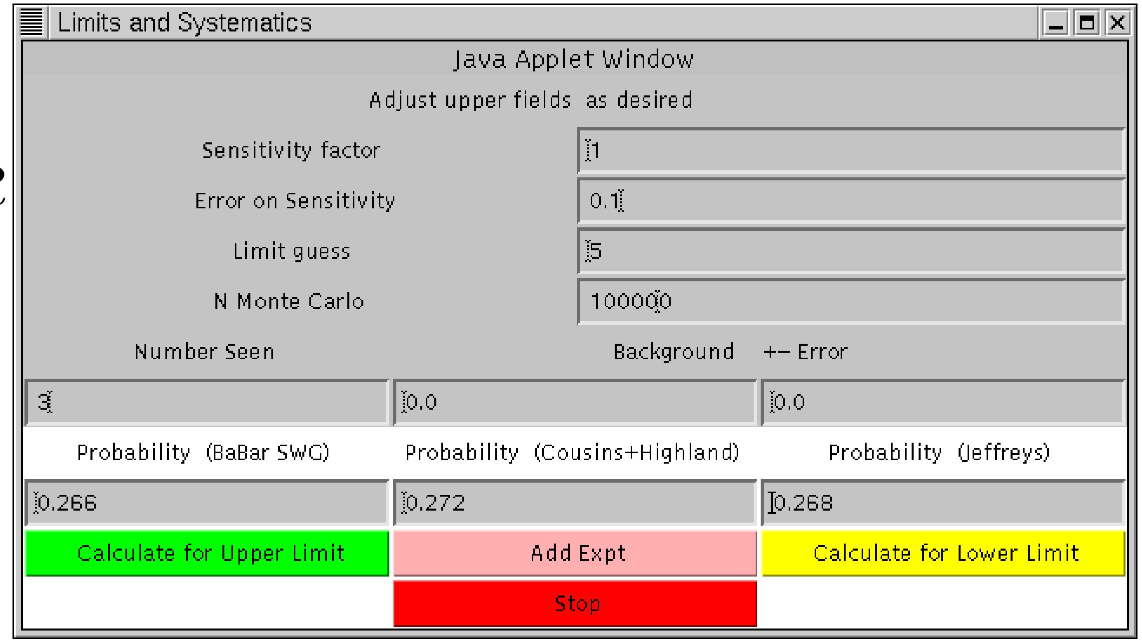
Limits on R from observed n.

Cousins and Highland: $n = SR$

BaBar: $R = An \qquad A \equiv 1/S$

Suppose you observe 3 events

Uncertainty of 10% on S or A



$P(\leq 3)$ from $R = 5.0$ is 27.2% (C & H) but 26.6% (BaBar SWG)

Why? Ambiguity in prior; Gaussian in S is not Gaussian in A.

Jeffreys' prior (uniform in lnA) gives intermediate result.

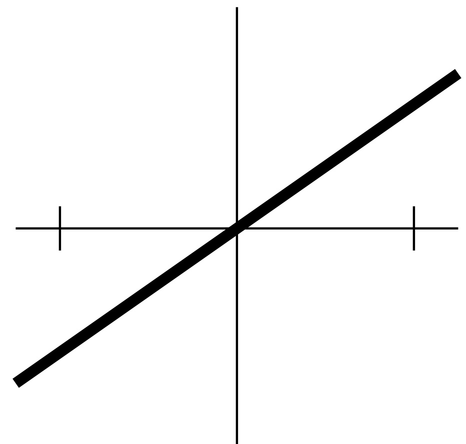
II: Evaluating effects of systematic uncertainty

Myth: This can't be done by standard technique

Combination of errors (with correlation terms) is fine for small changes

For large or non-differentiable changes, shift parameter by $\pm\sigma$ and re-evaluate result (or $\pm 2\sigma$ and re-evaluate and halve, or...) Can do this for final or intermediate result.

Example: Background evaluated from MC: depends on tuning parameter which has some uncertainty.



Can simplify, e.g. vary cuts rather than energy scale.

This is not a special procedure for systematic errors. It's just standard combination-of-errors.

Evaluation: the error on error paradox

Result R. Systematic effect a.

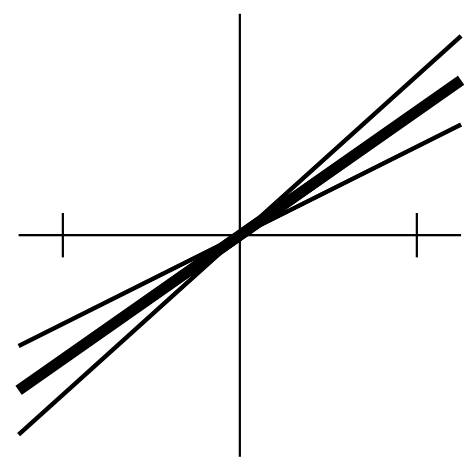
Quoted result is $R(a)$. Systematic uncertainty σ_a. Calculate $R(a-\sigma_a)$ and $R(a+\sigma_a)$ and thus $R' = \frac{dR}{da}$

Lots of MC data at central value of a; less at $a \pm \sigma_a$, thus R' has error $\sigma_{R'}$ due to MC statistics. What is the Systematic error on R?

1 $\sigma^2 = (R'\sigma_a)^2 + (\sigma_{R'}\sigma_a)^2$

2 $\sigma^2 = (R'\sigma_a)^2 - (\sigma_{R'}\sigma_a)^2$

3 $\sigma^2 = (R'\sigma_a)^2$



Justification for (1)

Add in quadrature as the uncertainty in R' is another uncertainty and uncertainties add in quadrature

Justification for (2)

This R' has been modified from true R' in such a way that $< R'^2 >$ is increased. E.g. if R is really independent of a, ($R' = 0$) MC statistics errors will force R' away from zero. Compensate.

Justification for (3)

No messing. If this is important demand more MC data

Simplified example

Integer x generated with uniform probability in a large range. You want the best value of x^2. But you are given $y = x \pm 1$ where $\pm$ has equal probability.

Argument 1(Bayesian): $y = 5$ could have come from $x = 6$ or $x = 4$ with equal probability. The answer looks like 25 but could be 16 or 36. Add 1 to compensate.

Argument 2(Frequentist): $x = 5$ could give $y = 4$ or $y = 6$ with equal probability. 25 becomes 16 or 36. On average this is 26, so subtract 1.

$$y = x \pm 1 \qquad x = y \mp 1$$

$$y^2 = x^2 \pm 2x + 1 \qquad x^2 = y^2 \mp 2y + 1$$

Test case zero

You observe $y = 0$.

Argument 1 gives $0^2 + 1 = 1$. Which is spot on. $x = -1$ or $+1$ so $x^2 = 1$.

Argument 2 gives $0^2 - 1 = -1$. Which looks crazy. 'Must be wrong.'

Counterattack. Suppose you generate $x = 0$. This will give $y^2 = 1$ so argument 2 is spot on and argument 1 is out by 2. Argument 1 will never give 0.

You can never know $x = 0$, but you can know $y = 0$.

Is the 'Uniform distribution' tenable?

Distribution (prior) in x (or R') cannot be uniform from $-\infty$ to $+\infty$. You would be surprised at very large values. So $y = 5$ is more likely to be an upward fluctuation than a downward one.

...Argument 1 dead

Thoughts and conclusion

This *is* a valid frequentist problem (even if σ_a is Bayesian): can in principle rerun MC evaluation many times.

Argument 2 is 'technically' correct. Gives the unbiassed estimate. But it means that measurements with $\sigma_{R'} > |R'|$ must contribute negatively to the systematic estimate, on the grounds this compensates for overestimation in other results (parallel universes?).

To be right in general you may have to do something manfestly wrong in an individual case.

If you have a whole lot of such corrections this may be arguable, but not if it's unique.

You'll never get it past the referee. Go for Argument 3.

III: Checks: Finding Mistakes

Omitting a systematic effect is a mistake. Need to identify all possible factors - including implicit ones. And check for other mistakes.

Think through the analysis and

– Phone a friend

– Ask the audience

Then do all the checks you can think of

- Separate data subsets
- Change cuts
- Change histogram bin size
- Change parametrisations (inc. order of polynomial)
- Change fit technique
- Look for impossibilities

Good practice:

Example:

'... consistency checks, including separation of the data 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 asymmetry found.'

- *from the BaBar sin*2β *measurement*

What is a significant difference?

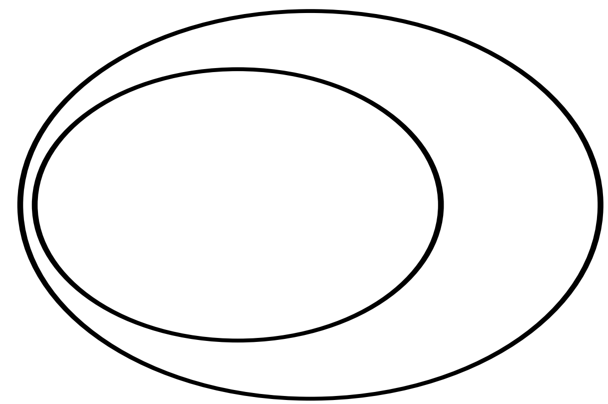
Standard analysis gives $a_1 \pm \sigma_1$. Check: different method gives $a_2 \pm \sigma_2$

Almost certainly $a_1 \neq a_2$! But want $\Delta = a_1 - a_2$ to be 'small'.

'Small' is not 'below the statistical error'. Analyses (may) share data

$\sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$

Suppose estimate is a mean and check uses a subset



$a_1 = \frac{1}{N_T}\sum_T x_i \qquad a_2 = \frac{1}{N_S}\sum_S x_i$

$\sigma_1 = \frac{\sigma}{\sqrt{N_T}} \qquad \sigma_2 = \frac{\sigma}{\sqrt{N_S}}$

$Cov(a_1, a_2) = N_S \frac{1}{N_T}\frac{1}{N_S}\sigma^2$

$\rho = \sigma_1/\sigma_2$

$\sigma_\Delta^2 = \sigma_2^2 - \sigma_1^2$: subtract in quadrature

General case

$\sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$

Introduce (briefly) $a(w) = wa_1 + (1-w)a_2$

Variance

$$\sigma_{a(w)}^2 = w^2\sigma_1^2 + (1-w)^2\sigma_2^2 + 2w(1-w)\rho\sigma_1\sigma_2$$

Choose w to get smallest variance

$$\frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

even this cannot be smaller than the Minimum Variance Bound σ_0

Translate to a limit on ρ and then a limit in σ_Δ gives

$$\sqrt{(\sigma_1^2 - \sigma_0^2)} + \sqrt{(\sigma_2^2 - \sigma_0^2)} \geq \sigma_\Delta \geq \left| \sqrt{(\sigma_1^2 - \sigma_0^2)} - \sqrt{(\sigma_2^2 - \sigma_0^2)} \right|$$

If $\sigma_1 = \sigma_0$ reduces to subtraction in quadrature again

Judgement day

Is there a problem? Make decision depending on size of discrepancy, number of checks being done, and basic plausibility.

If it passes do Nothing!

Do not add (small) discrepancy to systematic error.

1) It's silly

2) It penalises diligence

3) Errors get inflated. (LEP experiments agreed with each other and the Standard Model far *too* well.)

Be careful. Contrast moving mass cuts by defined amount to compensate for energy uncertainty (evaluation and included) and changing mass cuts by arbitrary amount to check efficiency/purity (consistency and not included if successful.)

What to do if it fails.

1: Check test. Find and fix mistake.

2: Check analysis. Find and fix mistake.

3: Worry. Maybe with hindsight an effect is reasonable. This check now becomes an evaluation.

4: Worry. It may only be the tip of the iceberg

Last resort: Incorporate in systematic error.

Illustration: inappropriate function

True calibration $y = x + 0.3x^2$

You fit $y = mx + c$

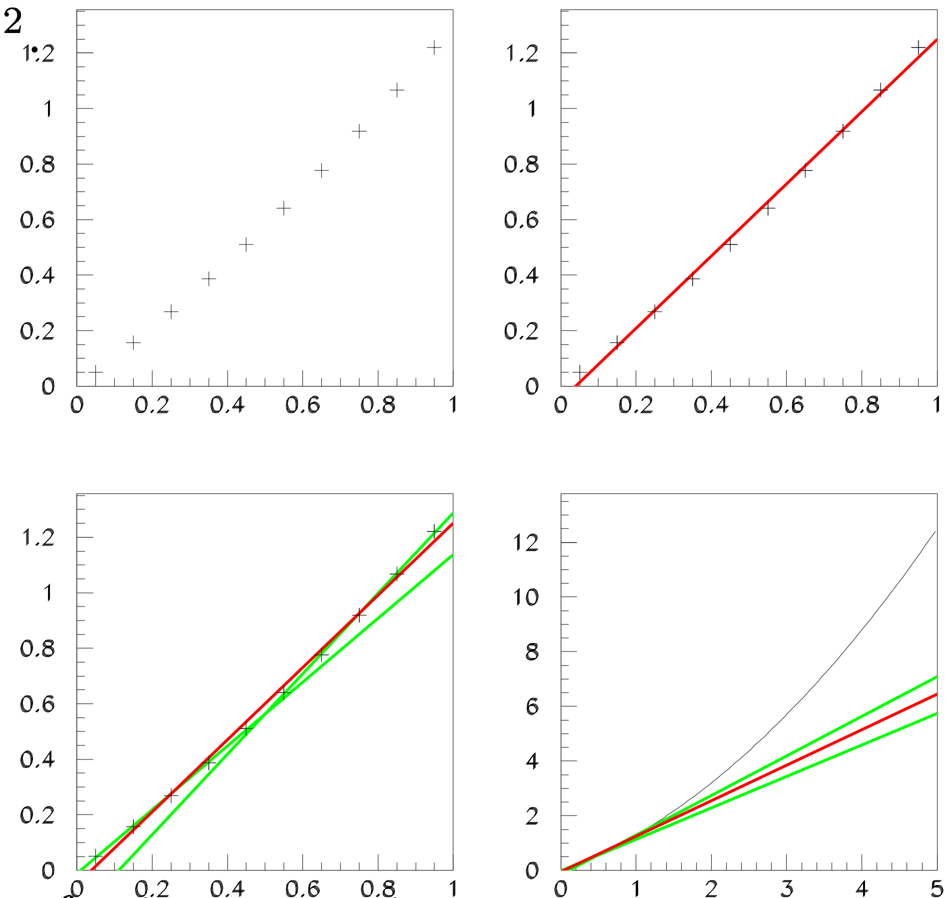
(get slope 1.3)

Check: calibrate subranges

Different slopes!

(1.15 and 1.45)

What to do?
Add calibration difference
to systematic error?
In range 0-1 far too harsh.
Outside 0-1 far too lenient.



There is no 'correct' procedure for incorporation.

Common Practice

How to write a paper/thesis

1: Devise cuts, get result

2: Do analysis for random errors (likelihood or Poisson statistics.)

3: Make big table

4: Alter cuts by arbitrary amounts, put in table

5: Repeat step 4 until time/money/supervisor's patience is exhausted

6: Add variations in quadrature

7: Quote result as 'systematic error'

8: If challenged, describe it as 'conservative'

This combines evaluation of errors with checks for mistakes, in a totally inappropriate way.

Conclusions: advice for practitioners

- Thou shalt never say ‘systematic error’ when thou meanest ‘systematic effect’.
- Thou shalt know at all times whether what thou performest is a check for a mistake or an evaluation of an uncertainty
- Thou shalt not incorporate successful check results into thy total systematic error and make thereby a shield behind which to hide thy dodgy result.
- Thou shalt not incorporate failed check results unless thou art truly at thy wits’ end
- 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.