Errors and Simulation

The mission

• given some measurements
• what are errors in final results?

Some analytical rules

Numerical (simulation) methods and Übung
General comments

• Vorlesungen
  • 2+1 oder
  • 2+ 2(jede zweite Woche)
• Folien in Stine und web

• Andrew Torda
  • schade@zbh.uni-hamburg.de (Mo-Do)
• Marco Matthies (Übungen)

• Sprache ?

• Much of material here from AST Kurs
Lehrbuch

• Diese Vorlesungen
  Berendsen, H.J.C., "A student’s guide to data and error analysis", Cambridge University Press 2011
• Normaleweise
  • Handouts / Primärliteratur

Languages

• Englisch / Deutsch ?
• C ? Perl ?
• ruby, python ?
• R ? Maple ?
Nomenclature

Numbers
• 2,137.33 Englisch
• 2.137,33 Deutsch
• 2 137.33 Wissenschaftlich

Units
• SI Einheiten
  • s not sec / min / hr
• Exceptions..
  • cm$^{-1}$, Å ...
Style

main rule
• use 1 style e.g.
  • $r$ scalar variables
  • $\mathbf{r}$ a vector
  • $\mathbf{R}$ a matrix

Formula
• $y = ax$ not $y = a \times x$ (computer language)

Verboten
• mixing fonts
• the variable $r$ is very important in $A = \pi r^2$
• call $\sigma$ the standard deviation as in $\sigma = f(\{x\})$
Conventions – no complaints

Σ is a sum
  • in other lectures, Σ is an alphabet
statistics books?
  • X might be a scalar
π can be a probability (physics)
Magical constants, conventions
  • $k, h, R, \ldots, t, T, \beta$
  • $pV = nRT$ horrible mix of upper and lower case
Errors and Simulation

Typical questions

A ⇌ B + C

I know $k_{eq}$, [A] and [C] with errors.
what is error in estimated [B]?

Enzyme kinetics

$v = \frac{v_{max}[S]}{k_m + [S]}$

how does error in $v$ depend on error in [S]?

To come

• analytical rules
• excuse to simulate
Distributions

Usually assume

- normal distributed errors
  \[ f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) \]
- random errors (not systematic)
- keep remembering \( \mu \), mean of distribution

\( \text{E}(a) \) "expectation value" of \( a \)
  - weighted average
Adding errors

• $f = x + y$  but we have errors
• given errors in $x$ and $y$, what is the error in $f$?

• define Standardabweichung (standard deviation) $\sigma$
• Variance $\equiv \sigma^2$
• $E(x)$ expected value of $x$
• standard deviation.. average of deviations from mean
• variance? $\sigma_x^2$

\[
\sigma_x^2 = E((x - \mu_x)^2) \quad \text{and} \quad \sigma_y^2 = E((y - \mu_y)^2)
\]

\[
\sigma_f^2 = E\left((f - \mu_f)^2\right)
\]
Problems

• How long is this room?
  • measured with a ruler with ± 0.5 cm accuracy

• Area of room?
  • measured with a ruler with ± 0.5 cm accuracy
Adding errors \( f = x + y \)

\[
\sigma_f^2 = E((f - \mu)^2) \\
= E\left((x + y - \mu_x - \mu_y)^2\right) \\
= E\left((x - \mu_x + y - \mu_y)^2\right) \\
= E\left((x - \mu_x)^2 + (y - \mu_y)^2 + 2(x - \mu_x)(y - \mu_y)\right) \\
= \sigma_x^2 + \sigma_y^2 + 2E(x - \mu_x)(y - \mu_y)
\]

- what do you expect for \( E(x - \mu_x)(y - \mu_y) \)?
  (\( x \) and \( y \) are uncorrelated)
- \( \sigma_f^2 = \sigma_x^2 + \sigma_y^2 \)
- \( \sigma_f = (\sigma_x^2 + \sigma_y^2)^{1/2} \) if you prefer standardabweichung
Many terms added (subtracted)

- $\sigma_f^2 = \sigma_x^2 + \sigma_y^2 + \cdots$

or

$\sigma_f^2 = \sum_i \sigma_{x_i}^2$
Multiplying errors? and more

• $f = xy$

• use relative uncertainties

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

• generalise this to other functions
• secret is in partial derivative
• contribution from each term is

• $\sigma_f^2 = \left| \frac{\partial f}{\partial x} \right|^2 \sigma_x^2 + \cdots$

• does this agree with previous versions?

Andrew Torda
18/12/2012
General adding of errors

general rule for everything everywhere

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

some prefer

\[ \sigma_f = \left( \left( \frac{\partial f}{\partial x} \right)^2 \sigma_x^2 + \left( \frac{\partial f}{\partial y} \right)^2 \sigma_y^2 + \ldots \right)^{1/2} \]

check if this agrees with previous claim
simple error

- $f = xy$ so $\frac{\partial f}{\partial x} = y$ and $\frac{\partial f}{\partial y} = x$

\[
\sigma^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2
\]

\[
= y^2 \sigma_x^2 + x^2 \sigma_y^2
\]

\[
\frac{\sigma^2}{x^2y^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}
\]

\[
\frac{\sigma^2}{f^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}
\]

which should agree
examples

- \( f = \ln x \) so \( \left( \frac{df}{dx} \right)^2 = \frac{1}{x^2} \)

- \( \sigma_f^2 = \left( \frac{df}{dx} \right)^2 \sigma_x^2 = \frac{\sigma_x^2}{x^2} \)

- \( \sigma_f = \frac{\sigma_x}{x} \)
division

• $f = \frac{x}{y}$ then $\frac{\partial f}{\partial x} = y^{-1}$ and $\frac{\partial f}{\partial y} = -xy^{-2}$

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

$$
= \sigma_x^2 y^{-2} + \sigma_y^2 x^2 y^{-4}
$$

$$
\frac{\sigma_f^2}{f^2} = \left( \sigma_x^2 y^{-2} + \sigma_y^2 x^2 y^{-4} \right) \frac{y^2}{x^2}
$$

$$
= \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2}
$$

• same as product rule
Example calculation

\[ A \rightleftharpoons B \quad \text{and} \quad k_{eq} = \frac{[B]}{[A]} \]

• measure \([A] = 10 \pm 0.5 \text{ M and } [B] = 1 \pm 0.05 \text{ M}\)

• \(k_{eq} = 0.1\)

• as before \(x = [B] \quad y = [A] \quad f = k_{eq}\)

\[
\frac{\sigma_f^2}{f^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} = \frac{(0.05)^2}{1} + \frac{(0.5)^2}{100} = 5 \times 10^{-3}
\]

• but we want the absolute error

\[
\sigma_f^2 = 5 \times 10^{-5} \quad f^2 = 10^{-2}
\]

\[
\sigma_f = 7.1 \times 10^{-3}
\]

• \(k_{eq} = 0.100 \pm 0.007 \quad \text{plausible?}\)
Plausible?

original errors $10 \pm 0.5 \text{ M}$ and $1 \pm 0.05$

- 5 %

final error

- $k_{eq} = 0.100 \pm 0.007$
- 7%

why do I write 0.100 and not 0.1?
Binding constant example

- Medikamente (Ligand) binds to protein
- \( P + L \rightleftharpoons PL \)  (Protein & Ligand)
- \( k_{eq} = \frac{[PL]}{[P][L]} \) what is a good drug? < nM\(^{-1}\) (10\(^{-9}\)M)
- can one measure \([P]\) ?\([PL]\) ?
- Must one measure \([L]\)? If we know \([L]_0\)
  - \([L] = [L]_0 - [PL]\)
  - \( k_{eq} = \frac{[PL]}{[P]([L]_0-[PL])}\)
- what to do?
\[ k_{eq} = \frac{[PL]}{[P]([L]_0 - [PL])} \]

looks like \( f = \frac{x}{y(c-x)} \)

Two possibilities

1. Have an excess of ligand then

\[ [L]_0 - [PL] \approx [L]_0 \]

so we have

\[ f = \frac{x}{yc} = \frac{1}{c} \frac{x}{y} \]

2. Treat properly

\[ \frac{\partial f}{\partial x} = \frac{c}{y(c-x)^2} = \frac{[L]_0}{[P]([L]_0 - [PL])^2} \]

and do the same for \( c, y \)

more examples in Übungzeit
Correlations

Problems – earlier:

- \[ \sigma_f^2 = E((f - \mu)^2) \]

\[
\ldots
\]

\[ = \sigma_x^2 + \sigma_y^2 + 2E(x - \mu_x)(y - \mu_y) \]

- claim that \( E(x - \mu_x)(y - \mu_y) = 0 \)

- implicit
  - \( \Delta x \) and \( \Delta y \) sometimes +, sometimes –

- what if they are correlated (not independent)?
More problems

• Non-independent variables is a problem
• worse
  • I suspect a systematic influence of $x$ on $y$
    • (some detector bias)
  • I have a complicated expression
    
    $$2[A] + [B] \rightleftharpoons [C] + [D]$$
    $$\Delta G = -RT \ln \frac{[C][D]}{[A]^2[B]}$$
    • measured volumes, masses, temperature, spectroscopic absorption, ...
    • 5 repetitions with different [C]
  • real example...
Calculation of protein structures

• 4 atoms in space with a dihedral angle
  • \( \tau \) is estimated based on a measurement from NMR (with errors)
  • what is \( \sigma_{d_{hk}} \)?
  • \( \theta_{hij} \) from standard tables, but fluctuates at room temperature

• \( d_{hk}^2 = \)
  \[ (d_{ij} - d_{hi} \cos \theta_{hij} - d_{jk} \cos \theta_{ijk})^2 + (d_{hi} \sin \theta_{hij} - d_{jk} \sin \theta_{ijk} \cos \tau_{hijk})^2 + (d_{jk} \sin \tau_{hijk})^2 \]
Simulations

• what to do when analytical versions are too hard

• Ingredients:
  • random number generator (normal distribution)
  • model / formula
  • ability to calculate $\bar{x}, \sigma$ from a set of numbers
  • histogram

• Discuss some steps
• pseudo code
random numbers

Good / bad
• repeat cycle – after \( n \) numbers, they repeat
• no detectable correlations, biases
  • number \( n \) must not be a predictor of \( n + 1 \)

Linear congruential
• most primitive, but useful – would be OK here
  \[ x_{n+1} = (ax_n + c) \mod m \]
  • magic values of \( a, c \) and \( m = 2^{32} \)
• uniformly distributed
Arbitrary distributions

Shifting to range \([a..b]\) (add \(a\) and scale by \(b - a\))

To get a distribution
- special tricks
  - Box-Mueller for gaussian
  - adding uniform numbers for a gaussian
- variable transform methods
  - OK – not always easy to apply
- rejection methods
  - slow
  - easy to apply
  - very widely applicable
For error estimations

Belief / model / simplification:
• there is a value $E(x)$, $\bar{x}$, ..
• we measure $x + \Delta x$ and $\Delta x$ is normal distributed

We want a normal distribution
rejection methods

Define rand() as function which returns numbers [0..1]

• to generate numbers from a distribution, \( p(x) \) from [min..max]

\[
\text{while (true)} \\
\quad r = \text{rand(min.. max)} \\
\quad \text{prob} = \text{dist}(r) \\
\quad \text{if prob} < \text{rand}(0..1)) \\
\quad \text{return } r
\]
simple simulation

• we have a relation like  \[ v = \frac{v_{max}[S]}{k_m+[S]} \]
• you can measure [S], \( v \) with 10 % and 20 % error
• \( k_m \) from the literature with 20 % error
• how accurately can you estimate \( v_{max} \)?
• \( v_{max} = \frac{v(k_m+[S])}{[S]} \)

• ingredients
  • gauss(\( \mu, \sigma \)) function
  • \( v_{max}(v, k_m, [S]) \) function
  • plotting and statistics
Simple simulation

get measurements of \( v, k_m, s \) and associated \( \sigma \)
for \( (i = 0; i < \text{MAX\_STEP}; i++) \) {
    \( v = \text{gauss}(v, \sigma_v) \)
    \( k = \text{gauss}(k_m, \sigma_{k_m}) \)
    \( s = \text{gauss}(s, \sigma_s) \)
    \( \text{tmp} = \text{vmax}(v, k, s) \)
    /* store \text{tmp}, calculate ...*/
}

Andrew Torda
18/12/2012
Why simulate ? Why do analytically ?

Analytical methods are usually better
• no numerical problems, convergence
• man braucht keinen Computer
• no new program for every problem
Why simulate?

- there were approximations used in analytical approach
- other error distributions
  - values lie between 1 and 2, not normally distributed
  - values from 0 to 1 have 10% error
    for $x > 1$, 30% error
  - non-gaussian errors
- easy to try out scenarios with plausible numbers
- correlations
  - $(x - \Delta x)(y - \Delta y)$ does not average to zero
  - spectroscopy: much $x$ affects signal from $y$
- non-symmetric distributions of final errors..
non-symmetric distributions

- mean ≠ most common value
- errors are not symmetric
- important?
  - particle velocities
  - sequence alignment scores
  - kinetics results

Berendsen, H.J.C., "A student’s guide to data and error analysis", Cambridge University Press 2011
Summarise

Errors
• additive, multiplicative, most general

No mistakes in nomenclature

Klausur
• general formula, ability to apply to new formula
• examples from Übung