Errors and Simulation

- The mission
  - given some measurements
  - what are errors in final results ?

- Some analytical rules
- Numerical (simulation) methods and Übung

Andrew Torda wintersemester 2011/2012

General comments

- Vorlesungen
  - Mi + Do 08:30
- 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

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

- verboten
  - mixing fonts
  - the variable r is very important in $A = \pi r^2$
  - call σ 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
- $A \rightleftharpoons B + C$
  - 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 μ, mean of distribution
- $\mathrm{E}(a)$ "expectation value" of a
  - weighted average

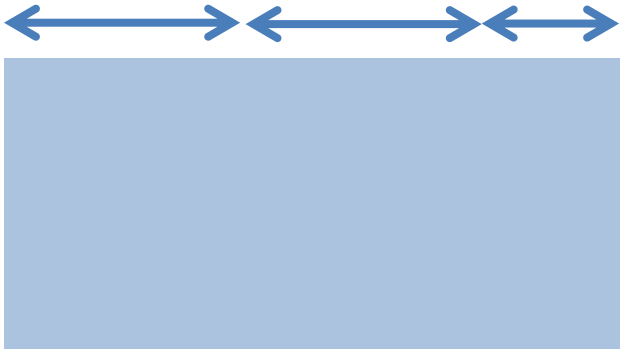
Adding errors

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

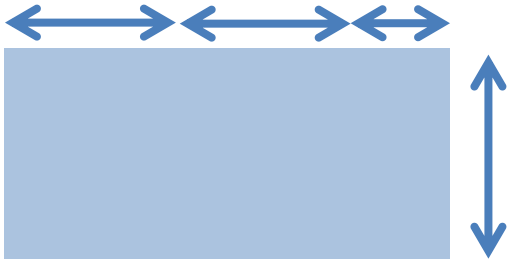
- define Standardabweichung (standard deviation) σ
- Variance $\equiv \sigma^2$
- $\mathrm{E}(x)$ expected value of x
- standard deviation.. average of deviations from mean
- variance ? σ_x^2
  - $\sigma_x^2 = \mathrm{E}\left((x - \mu_x)^2\right)$ and $\sigma_y^2 = \mathrm{E}\left(\left(y - \mu_y\right)^2\right)$
  - $\sigma_f^2 = \mathrm{E}\left(\left(f - \mu_f\right)^2\right)$

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

$$
\begin{aligned}
\sigma_f^2 &= E\left((f-\mu)^2\right) \\
&= E\left(\left(x+y-\mu_x-\mu_y\right)^2\right) \\
&= E\left(\left(x-\mu_x+y-\mu_y\right)^2\right) \\
&= E\left((x-\mu_x)^2+\left(y-\mu_y\right)^2+2(x-\mu_x)\left(y-\mu_y\right)\right) \\
&= \sigma_x^2+\sigma_y^2+2E(x-\mu_x)\left(y-\mu_y\right)
\end{aligned}
$$

- what do you expect for $\mathrm{E}(x-\mu_x)\left(y-\mu_y\right)$?

  (x and y are uncorrelated)
- $\sigma_f^2 = \sigma_x^2 + \sigma_y^2$
- $\sigma_f = \left(\sigma_x^2+\sigma_y^2\right)^{1/2}$ 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

- $f = xy$
- adding uncertainties
- 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 ?

General adding of errors

- general rule for everything everywhere

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

some prefer

- $\sigma_f = \left(\left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \cdots\right)^{1/2}$
- 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

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

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

- $\sigma_f = \frac{\sigma_x}{x}$

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

- $\mathrm{A} \rightleftharpoons \mathrm{B}$ and $k_{eq} = \frac{[\mathrm{B}]}{[\mathrm{A}]}$
- measure [A] =10 ±0.5 M and [B]=1±0.05 M
- $k_{eq} = 0.1$
- as before, $x = [\mathrm{B}]$ and $y = [\mathrm{A}]$ and $f = k_{eq}$
- $\frac{\sigma_f^2}{f^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} = \frac{(0.05)^2}{1} + \frac{(0.5)^2}{100} = 5 \times 10^{-3}$
- but we want the absolute error
- $\sigma_f^2 = 5 \times 10^{-3}\, f^2 = 5 \times 10^{-5}$
- $\sigma_f = 7.1 \times 10^{-3}$

- $k_{eq} = 0.100 \pm 0.007$ plausible ?

Plausible ?

- original errors 10 ±0.5 M and 1±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 $[L]_0$
  - $[L] = [L]_0 - [PL]$
  - $k_{eq} = \frac{[PL]}{[P]\left([L]_0 - [PL]\right)}$
- what to do ?

- $k_{eq} = \frac{[\mathrm{PL}]}{[\mathrm{P}]\left([\mathrm{L}]_0 - [\mathrm{PL}]\right)}$ looks like $f = \frac{x}{y(c-x)}$

- two possibilities
  1. Have an excess of ligand then

$$[\mathrm{L}]_0 - [\mathrm{PL}] \approx [\mathrm{L}]_0$$

so we have $f = \frac{x}{yc} = \frac{1}{c}\,\frac{x}{y}$

- 2. Treat properly $\frac{\partial f}{\partial x} = \frac{c}{y(c-x)^2} = \frac{[\mathrm{L}]_0}{[\mathrm{P}]\left([\mathrm{L}]_0 - [\mathrm{PL}]\right)^2}$
  - and do the same for c, y
- more examples in Übungzeit

Correlations

Problems – earlier:

- $\sigma_f^2 = E\left((f-\mu)^2\right)$

  $\ldots$

  $= \sigma_x^2 + \sigma_y^2 + 2\mathrm{E}(x-\mu_x)\left(y-\mu_y\right)$
- claim that $\mathrm{E}(x-\mu_x)\left(y-\mu_y\right) = 0$
- implicit
  - Δx and Δy 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

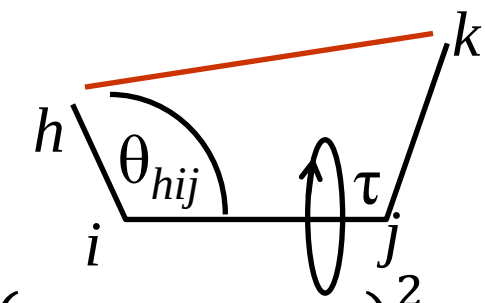
$$2[\mathrm{A}] + [\mathrm{B}] \rightleftharpoons [\mathrm{C}] + [\mathrm{D}]$$

$$\Delta G = -RT \ln \frac{[\mathrm{C}][\mathrm{D}]}{[\mathrm{A}]^2[\mathrm{B}]}$$

    - 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
  - τ is estimated based on a measurement from NMR (with errors)
  - what is $\sigma_{d_{hk}}$?
  - θ_{hij} from standard tables, but fluctuates at room temperature
- $d_{hk}^2 =$
  $\left(d_{ij} - d_{hi}\cos\theta_{hij} - d_{jk}\cos\theta_{ijk}\right)^2 +$
  $\left(d_{hi}\sin\theta_{hij} - d_{jk}\sin\theta_{ijk}\cos\tau_{hijk}\right)^2 + \left(d_{jk}\sin\tau_{hijk}\right)^2$



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
  - $x_{n+1} = (ax_n + c) \bmod m$
    - magic values of a, c and $m = 2^{32}$

  - uniformly distributed

Arbitrary distributions

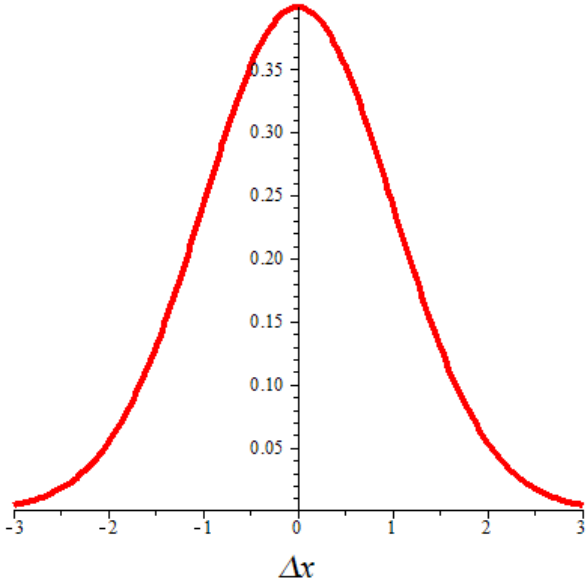
- shifting to range $[a..b]$ (add a and scale by $b - a$)

- 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

- our belief / model / simplification:
  - there is a value $\mathrm{E}(x)$, $\bar{x}$, ..
  - we measure $x + \Delta x$ and Δx is normal distributed

- we want a normal distribution

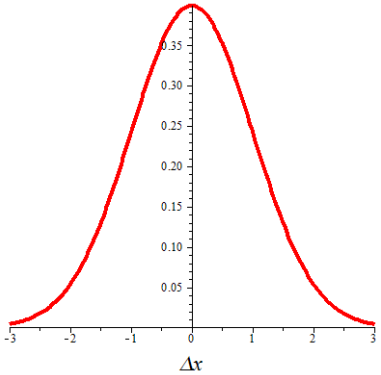


rejection methods

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

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

```
while (true)
        r = rand(min.. max)
        prob = dist(r)
        if prob < rand(0..1))
                return r
```



simple simulation

- we have a relation like $v = \frac{v_{max}[\mathrm{S}]}{k_m + [\mathrm{S}]}$
- 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} ?
- $v_{max} = \frac{v(k_m + [\mathrm{S}])}{[\mathrm{S}]}$

- ingredients
  - $\mathrm{gauss}(\mu, \sigma)$ function
  - $\mathrm{vmax}(v, k_m, [\mathrm{S}])$ function
  - plotting and statistics

Simple simulation

get measurements of v, k_m, s and associated σ

```
for (i = 0; i < MAX_STEP; i++) {
        v = gauss (v, σ_v)
        k = gauss (k_m, σ_{k_m})
        s = gauss (s, σ_s)
        tmp = vmax (v, k, s)
        /* store tmp, calculate …*/
}
```

Why simulate ?

Analytical methods are usually better

- no numerical problems, convergence
- man braucht keinen Computer
- no new program for every problem

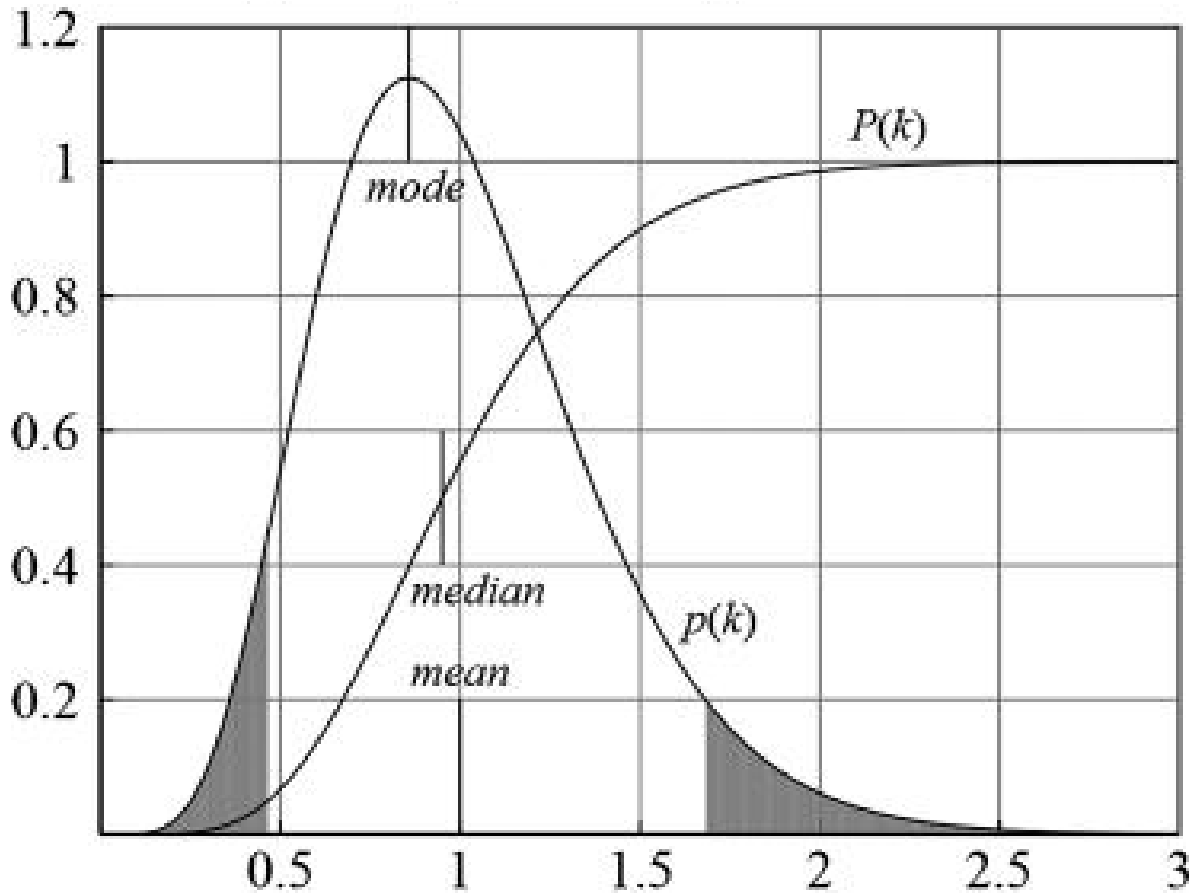
Why simulate ?

- 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