MASARYKOVA UNIVERZITA
INNOVATION LECTURES (INNOIEC)
www.muni.cz

Binding and Kinetics for Experimental Biologists
Lecture 7

Dealing with uncertainty: Confidence intervals

Petr Kuzmič, Ph.D.
BioKin, Ltd.

WATERTOWN, MASSACHUSETTS, U.S.A.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

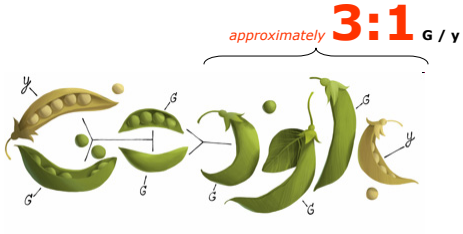
BioKin L t d.

"Hunches and intuitive impressions are essential for getting the work started, but it is only through the **quality of the numbers** at the end that the **truth** can be told."

-Lewis Thomas

L. Thomas (1977) "Biostatistics in Medicine", *Science* **198**, 675

Gregor **Mendel** (1822-1884)



Google - July 20, 2011

But how much **confidence** can you have in that number?

Lecture outline

- The problem:

  How much (or how little) can we trust our rate and equilibrium constants?

- The solution:

  Always report at least *some* measure of parameter uncertainty:

  - formal standard error
  - confidence interval
    (a) by systematic search (profile-*t* method)
    (b) by stochastic simulations (Monte-Carlo method)

- An implementation:

  Software *DynaFit*.

- An example:

  The classic "Biological oxygen demand (BOD)" problem

Part 1: Confidence intervals by systematic searching

"Profile-*t*" method

Example problem: Biological oxygen demand (B.O.D.)

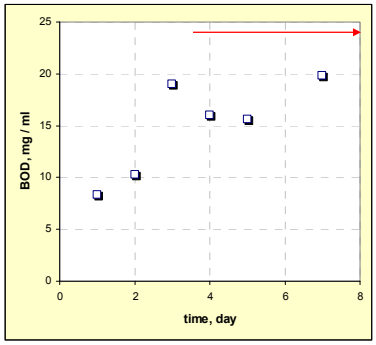
A CLASSIC DATA SET IN STATISTICAL LITERATURE

BOD = measure of **organic pollution** in environmental water

BOD at $t \rightarrow$ infinity ?

Table A1.4 Biochemical oxygen demand versus time.

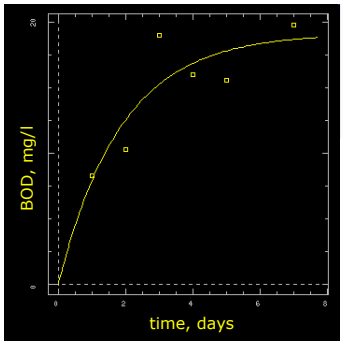
Time (days)	Biochemical Oxygen Demand (mg/l)	Time (days)	Biochemical Oxygen Demand (mg/l)
1	8.3	4	16.0
2	10.3	5	15.6
3	19.0	7	19.8



Bates D. M. & Watts, D. G. (1988)
Nonlinear Regression and its Applications
Wiley, New York, p. 270

Theoretical model: Exponential growth

COMPARE ALGEBRAIC MODEL WITH DYNAFIT NOTATION



Optimized Parameters

No.	Par#Set	Initial	Final
#1	k	1	0.531091
#2	[Oxygen]	10	19.1426

BOD_{max} = 19.1 mg/l

ALGEBRAIC MODEL:

$$B = B_{\max}\left[1 - \exp(-k\,t)\right]$$

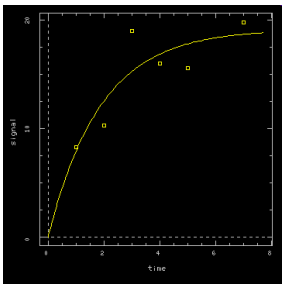
DYNAFIT MODEL:

```
[mechanism]

   Oxygen ---> Bacteria    :    k
```

How much should we trust these model parameters?

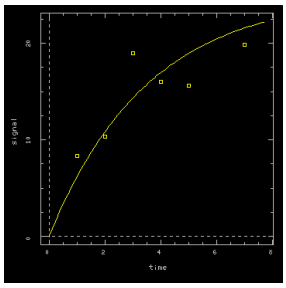
FIX B_{max} AT AN ARBITRARY VALUE, OPTIMIZE k

optimized parameter
fixed parameter



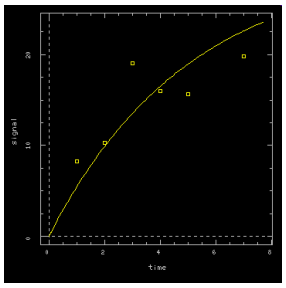
B_{max} = **19.1**
k = 0.53

sum of squares = **26.0**



B_{max} = **25.0**
k = 0.28

sum of quares = **41.9**

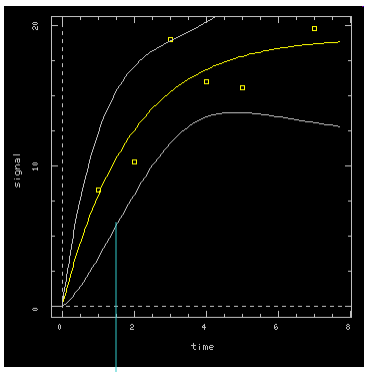


B_{max} = **30.0**
k = 0.20

sum of squares = **57.6**

A little better than nothing: Formal standard errors

THIS IS WHAT MOST PAPERS REPORT IN THE LITERATURE



Optimized Parameters

No.	Par#Set	Initial	Final	Std. Error
#1	k	1	**0.531091**	0.203082
#2	[Oxygen]	10	**19.1426**	2.49592

BOD_{max} = (19.1 ± 2.5) mg/l

implies the interval

19.1 – 2.5 = **16.6**
19.1 + 2.5 = **21.6**

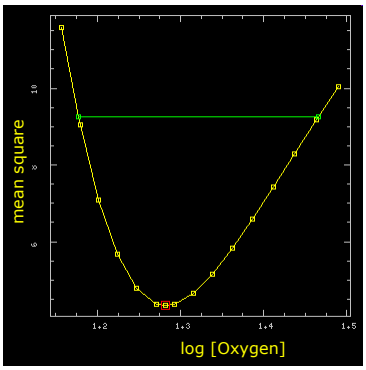
```
[settings]
{Output}
   InferenceBands = y
```

The correct way to do it: Approximate confidence intervals

VERY RARELY REPORTED IN THE LITERATURE (UNFORTUNATELY)

Optimized Parameters

No.	Par#Set	Initial	Final	Std. Error	CV (%)	Low	Low P (%)	High	High P (%)
#1	k	1	0.531091	0.203082	38.24	0.203907	90	1.2664	90
#2	[Oxygen]	10	19.1426	2.49592	13.04	15.0333	90	29.2602	90



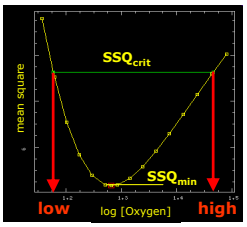
BOD_{max} = (19.1 ± 2.5) [15.0 – 29.3] mg/l

Confidence intervals: *Profile-t* method in DynaFit

A SEQUENCE OF SEVERAL INDEPENDENT LEAST-SQUARES FITS

INPUT:

```
[mechanism]
   Oxygen ---> Bacteria    :    k

[constants]
   k = 1 ?

[concentrations]
   Oxygen  = 10 ??
```



ALGORITHM:

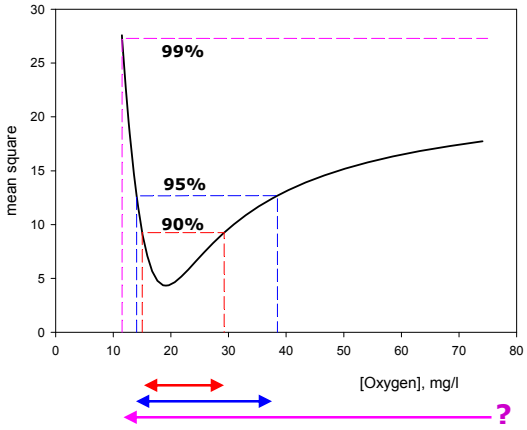
1. Perform an **initial fit** with all parameters optimized
2. Perform a series of **follow-up** fits focusing on a given parameter

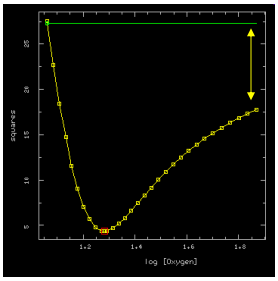
   2a. "Freeze" the parameter at values progressively further away from optimal
   2b. Optimize all remaining parameters
   2c. Repeat (2a) and (2b) until sum of squares reaches a "critical value" above minimum

REFERENCE: Bates, D. M., and Watts, D. G. (1988)
Nonlinear Regression Analysis and its Applications
Wiley, New York, pp. 127-130

Confidence level (%) and the width of confidence intervals

HIGHER CONFIDENCE LEVEL = WIDER CONFIDENCE INTERVAL



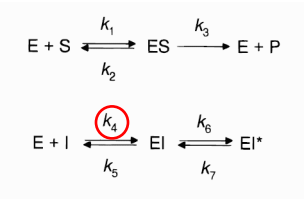


Upper limit for BOD_{max} *could not be determined* at 99% confidence level.

Example of a half-open confidence interval

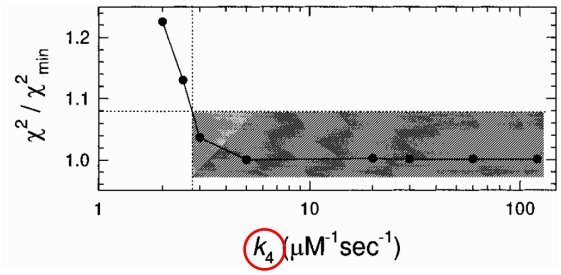
UPPER LIMITS FOR **BIMOLECULAR ASSOCIATION** RATE CONSTANTS OFTEN CANNOT BE DETERMINED

MECHANISM



$k_4 = (5 \pm 200)\ [3 - \infty]\ \mu M^{-1}s^{-1}$

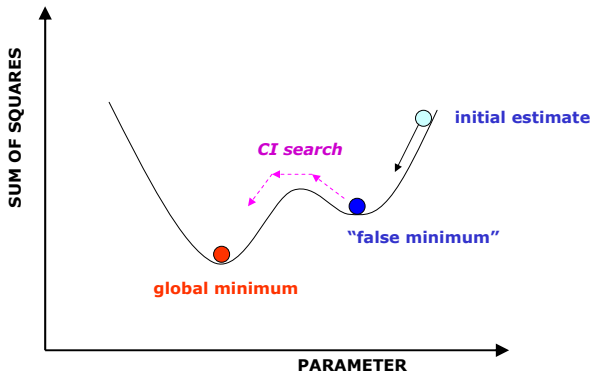
CONFIDENCE INTERVAL FOR k_4



Moss, Kuzmic, *et al.* (1996) *Biochemistry* **35**, 3457-3464.

Search for confidence intervals may diagnose "false minima"

AN OCCASIONAL SIDE-BENEFIT OF CONFIDENCE INTERVAL SEARCHES



SUMMARY: Confidence intervals via *profile-t* method

- Confidence intervals are **asymmetrical** for all *nonlinear* parameters
- Frequently much **wider** (more realistic) than ± formal standard errors
- Sometimes **half-open** intervals: "better than nothing", e.g. for bimolecular association
- Can have **mechanistic** implications (reversible / irreversible steps)
- Sometimes CI search helps in falling out of **false minima**
- In DynaFit scripts, CIs are requested by the **"??"** syntax
- Should always be reported with their corresponding **confidence levels** (%)
- CIs are wider at higher confidence levels
- Frequently used confidence levels: **90%, 95%, or 99%**
- Computation can be time consuming (many repeated least-squares fits)

```
[settings]

{Marquardt}
   ConfidenceLevel = 90
```

Part 2: Confidence intervals by stochastic simulations

Monte-Carlo method

Monte-Carlo confidence intervals: Algorithm

1. Perform an **initial fit** as usual
2. Perform a **large** series (> 1000) of **follow-up** fits

   2a. **Simulate** an artificial data set with **random errors** superimposed in ideal data
   2b. Perform a fit of the artificial data
   2c. Compile a **histogram** of distribution for model parameters from many repeated fits
   2d. Determine the range of plausible values for model parameters from the histograms

REFERENCE:

Straume, M., and Johnson, M. L. (1992)

"Monte-Carlo method for determining complete confidence probability distributions of estimated model parameters"

Methods Enzymol. **210**, 117–129.

Monte-Carlo confidence intervals: DynaFit input

A SINGLE LINE ADDED TO THE DYNAFIT SCRIPT

```
[task]
   task       = fit
   data       = progress
   confidence = monte-carlo

[mechanism]
   Oxygen ---> Bacteria    :    k

[constants]
   k = 1 ??

[concentrations]
   Oxygen  = 10 ??

...
```

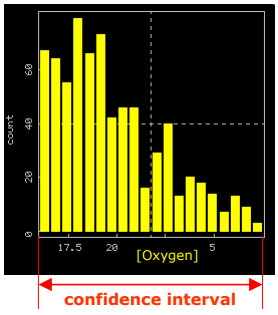
```
[settings]

{MonteCarlo}

   Runs                              = 1000
...
   ConcentrationErrorPercent         = 0
```

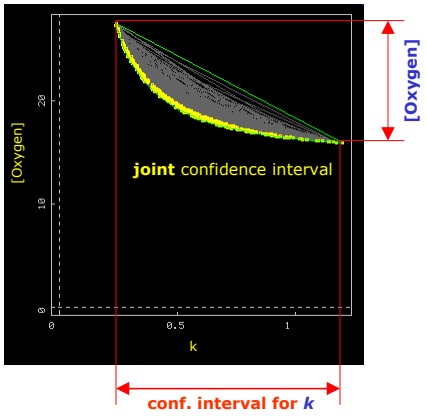
plus a number of other advanced control parameters

Monte-Carlo confidence intervals: DynaFit output

HISTOGRAMS OF DISTRIBUTION *PLUS* **CORRELATION** PLOTS

Distribution of best-fit values from 1000 least-squares fits of simulated data

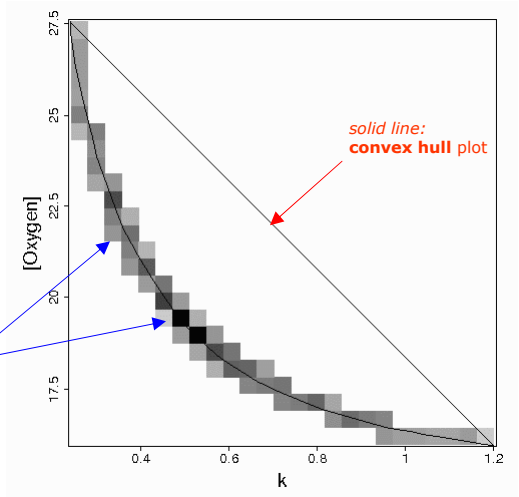




Monte-Carlo confidence intervals: Convex hull plots

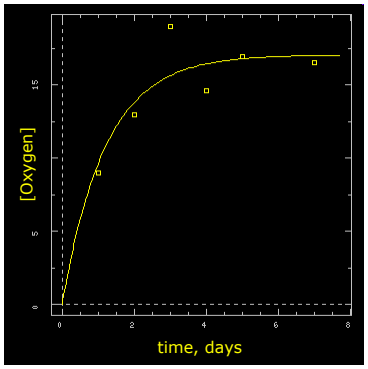
CONVEX HULL = SHORTEST PATH COMPLETELY ENCLOSING A GROUP OF POINTS IN A PLANE

EPS (PostScript) file generated by DynaFit:



Monte-Carlo and *profile-t* confidence intervals compared

MONTE-CARLO INTERVALS ARE ALMOST ALWAYS **NARROWER** THAN PROFILE-*t* AT 90% LEVEL



MONTE-CARLO METHOD (*n* = 1000)

	low	high
k	0.24	1.20
B_{max}	**16.0**	**27.5**

PROFILE-*t* METHOD (90% confidence level)

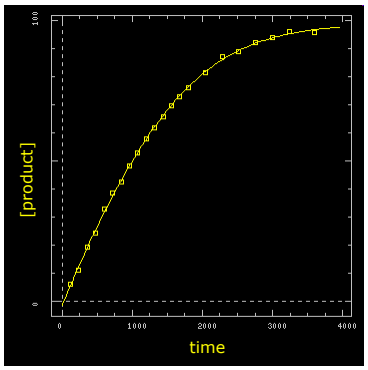
	low	high
k	0.20	1.27
B_{max}	**15.0**	**29.3**

good agreement between the two methods

Randomly varied concentrations: DynaFit input

REAGENT CONCENTRATIONS ARE ALWAYS AFFECTED BY RANDOM **TITRATION ERRORS!**

Enzyme kinetics: Substrate conversion

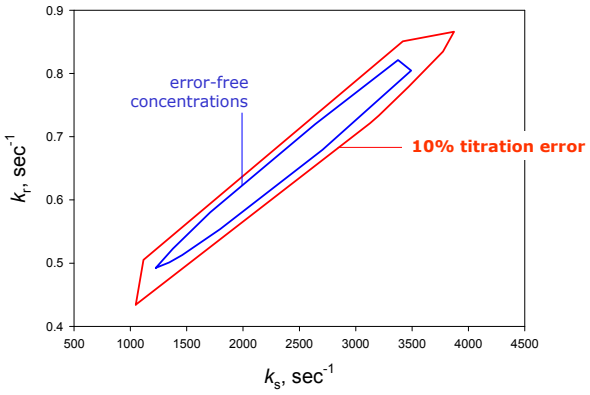


Mechanism: Michaelis-Menten

```
[mechanism]

   E + S <===> ES       :     k     ks
   ES ----> E + P       :     kr

[constants]

   k  =  100
   ks = 1000 ?
   kr =    1 ?

...

[settings]

{MonteCarlo}
   ConcentrationErrorPercent = 10

[end]
```


Randomly varied concentrations: DynaFit output

JOINT CONFIDENCE INTERVAL (AS A CONVEX HULL)



SUMMARY: Confidence intervals via *Monte-Carlo* method

PROS:

- Method makes **no assumptions** about the statistical **distribution** of model parameter errors
- Often uncovers "strange" effects such as **half-open** confidence intervals
  - mechanistic implications (**reversible / irreversible** steps)
- Reveals special patterns in the statistical **correlation** between model parameters
- Does not require an arbitrary choice of **confidence levels (%)**

CONS:

- Method makes **heavy assumptions** about the statistical **distribution** of **experimental errors**
  - could be overcome by the "shuffling" and "shifting" methods in DynaFit
- Can take a **very long time** to compute (multiple hours)
- Does **not** help in discovering false minima

Side comment: The issue of significant digits

Example of poor reporting: Hyperbolic fit in a student project

RESULTS FROM A SEMESTER-LONG RESEARCH PROJECT

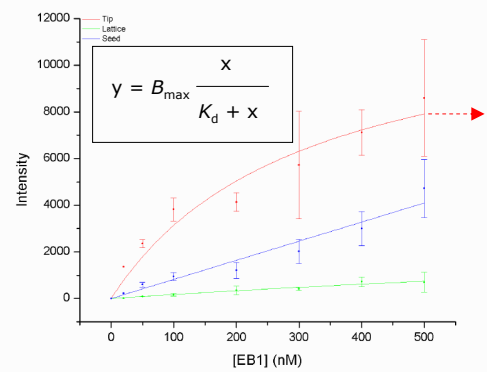


Table 1: Kd values assigned to EB1, EB2 and EB3 at each microtubule configuration

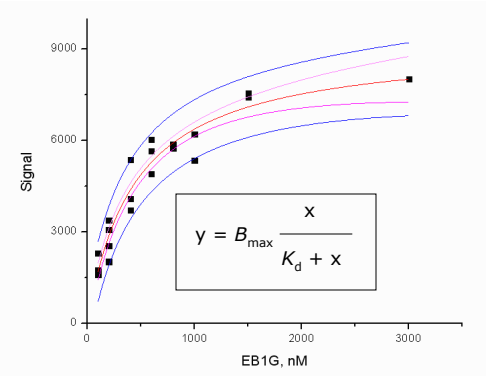
	Tip Kd	Lattice Kd	Seed Kd
EB1	307.09	N/D	N/D

what is wrong with this result?

1. no measure of uncertainty
2. too many digits

Software programs usually report too many digits

OUTPUT GENERATED BY **SOFTWARE PACKAGE "ORIGIN"**



Parameters

		Value	Standard Error
Signal	P1	9184.95075	542.54918
	P2	442.3346	67.39583

← B_{max}
← K_d

DIRECT OUTPUT FROM SOFTWARE:

K_d = (442.3346 ± 67.39583) nM

what is wrong with *this* result?

SENSIBLE WAY TO REPORT IT:

K_d = (440 ± 70) nM

RECIPE:

1. Round standard error to a **single significant digit**
2. Round best-fit value to the **same number of decimal points**

Overall summary and conclusions

ANY NUMERICAL RESULT REPORTED WITHOUT SOME MEASURE OF UNCERTAINTY IS **MEANINGLESS**

1. Always report at least some measure of statistical uncertainty
   for all nonlinear model parameters (rate and equilibrium constants).
2. At the very least report the formal ± standard errors.
3. Confidence intervals are more informative than standard errors.
4. DynaFit offers *two* different methods for confidence intervals:

   a. Systematic search (profile-*t* method)
   b. Stochastic simulation (Monte-Carlo method)
5. The two methods have their own merits and drawbacks
   When in doubt, use both.
6. DynaFit is not a "silver bullet": You must still use your brain a lot.