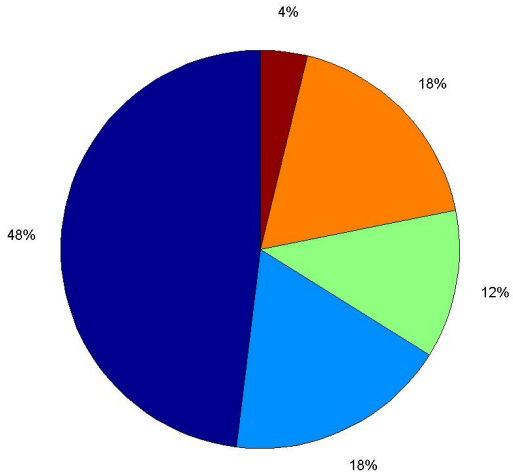
Introduction to Bayesian inference

M. L. Delignette-Muller - VetAgro Sup - LBBE

SETAC 2013, 05/12/2013

What is a probability ?

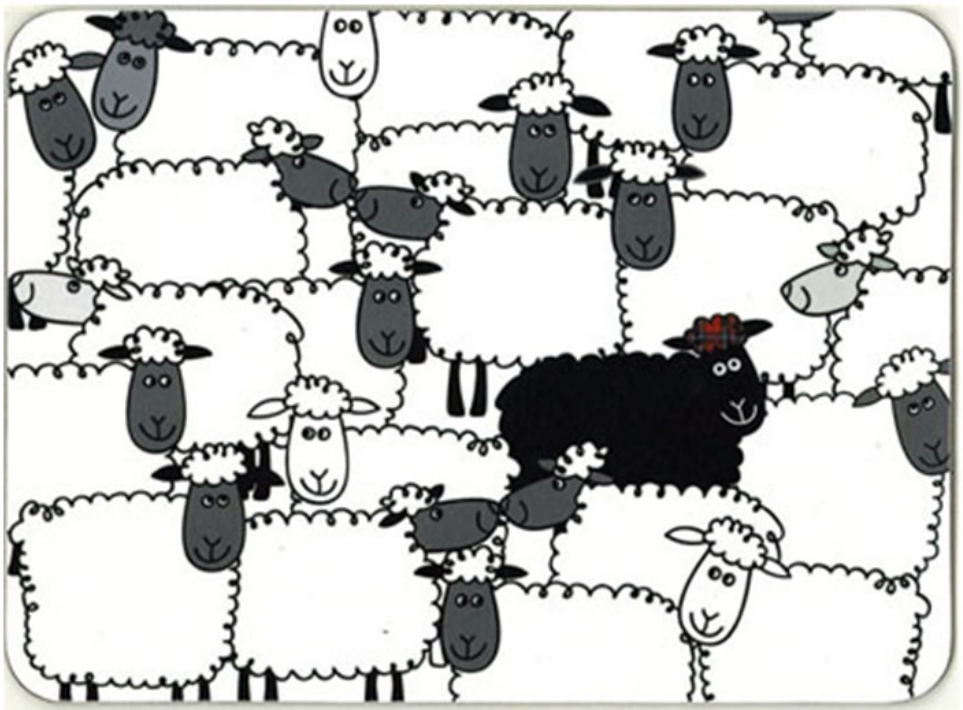


One word,
at least two definitions.

Frequentist view of probability

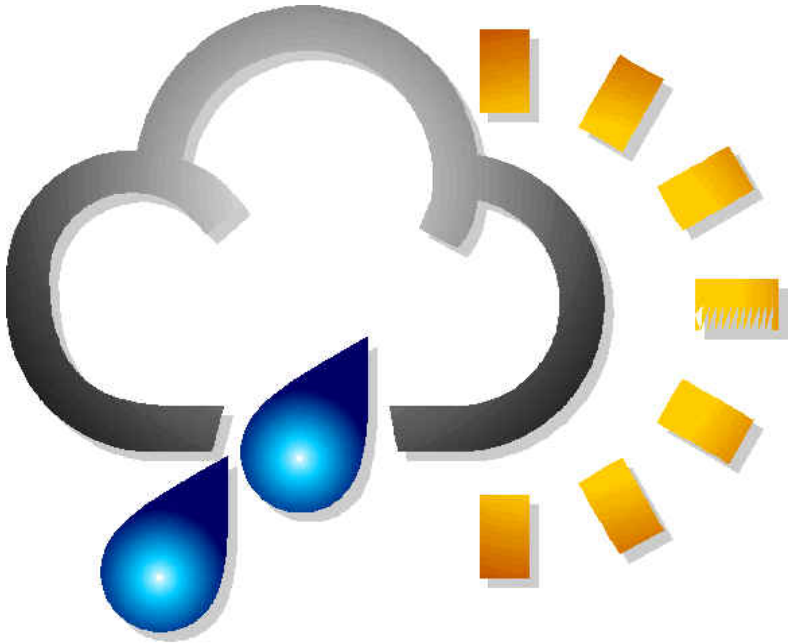
In a frequentist perspective, the probability of an event is defined as the fraction of times that the event occurs in a very large number of trials.

Probability of being a black sheep ?



Bayesian view of probability

In a bayesian prespective, the probability is seen as a degree of belief, a measure of uncertainty.



Probability of rain tomorrow ?

What is a model ?

A model

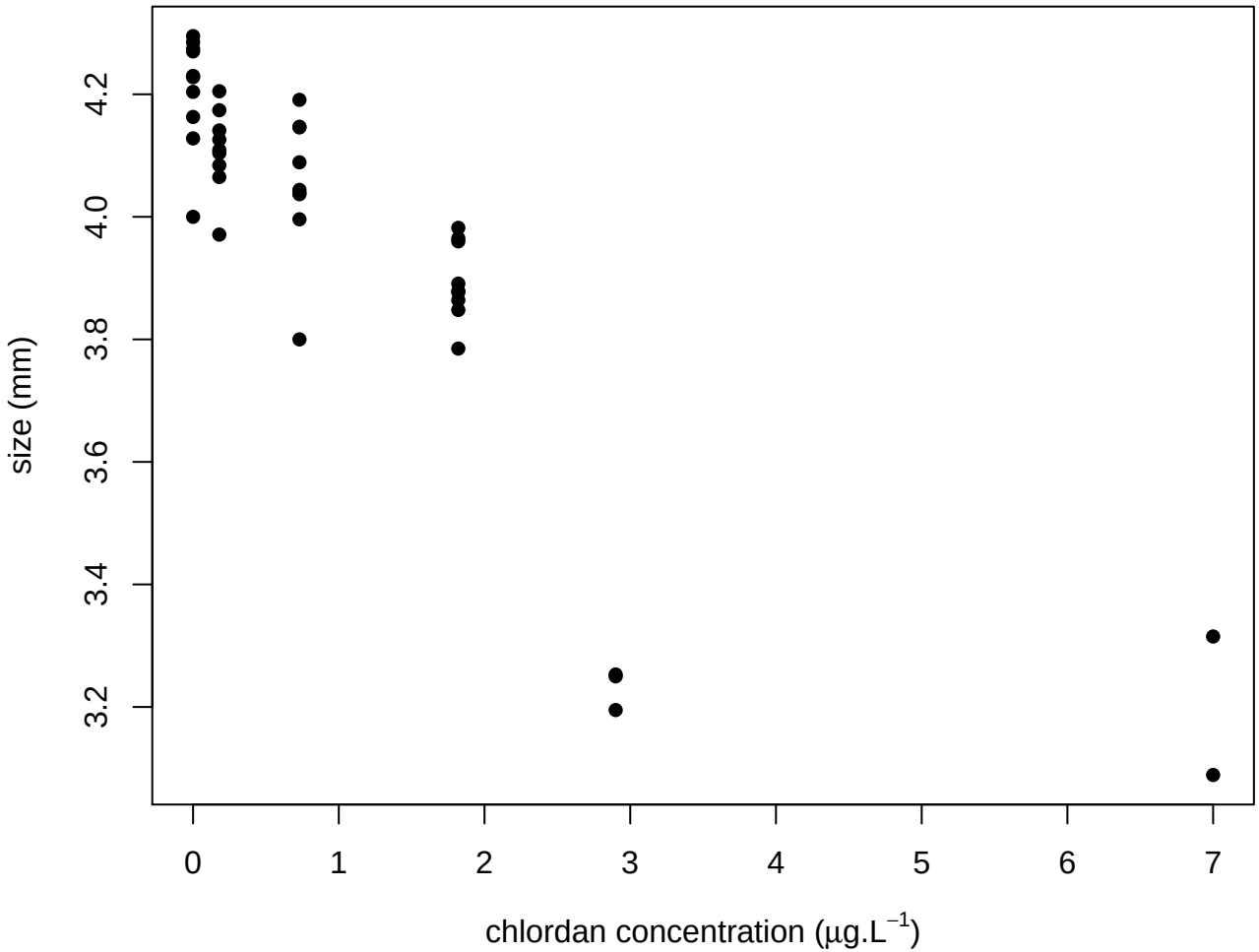
relates observed data Y
to a set of unknown parameters θ,
with the possible inclusion of fixed, known covariates X.
It is classically divided in two parts,

- the deterministic one $M(X, \theta)$,
- and the stochastic one (error model)

Example : the gaussian regression model
$Y = M(X, \theta) + \epsilon$ with $\epsilon \sim N(0, \sigma)$
A model may also be seen as a data generating process.

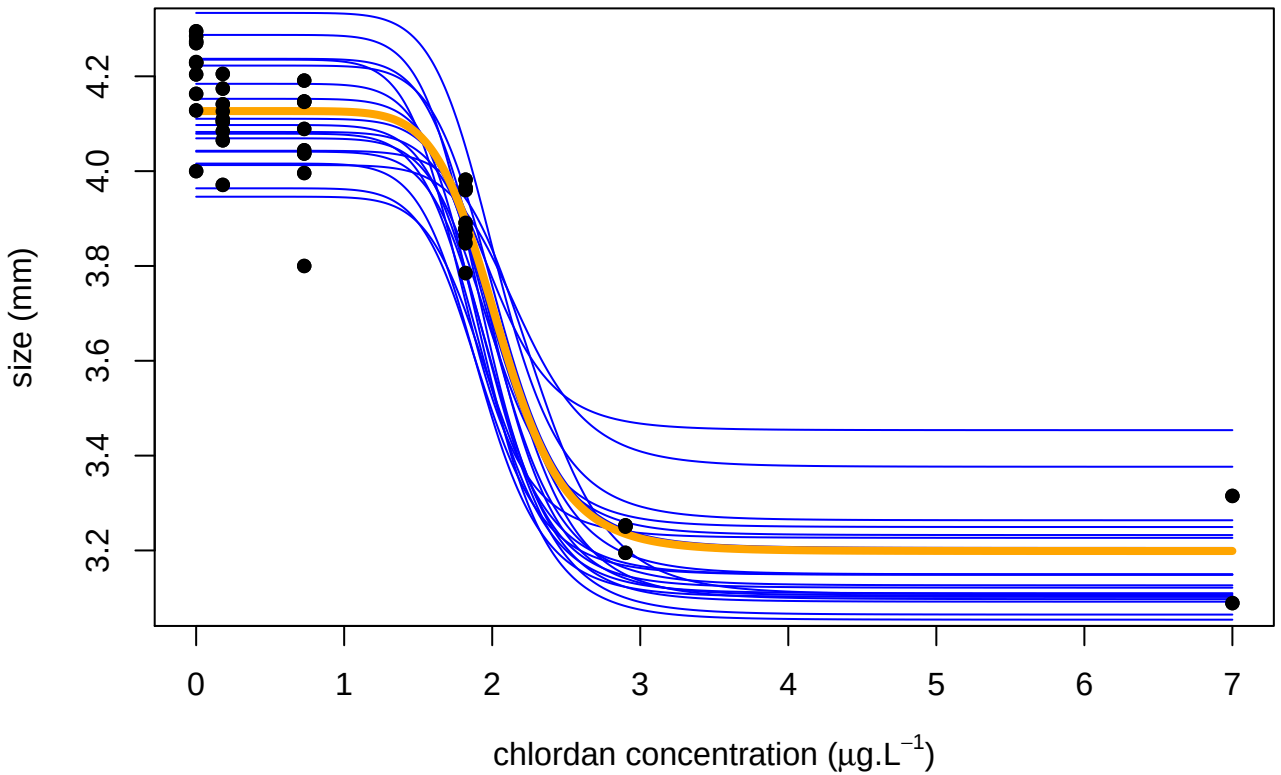
What is inference ?

Inference generally implies the fit of the model to observed data.



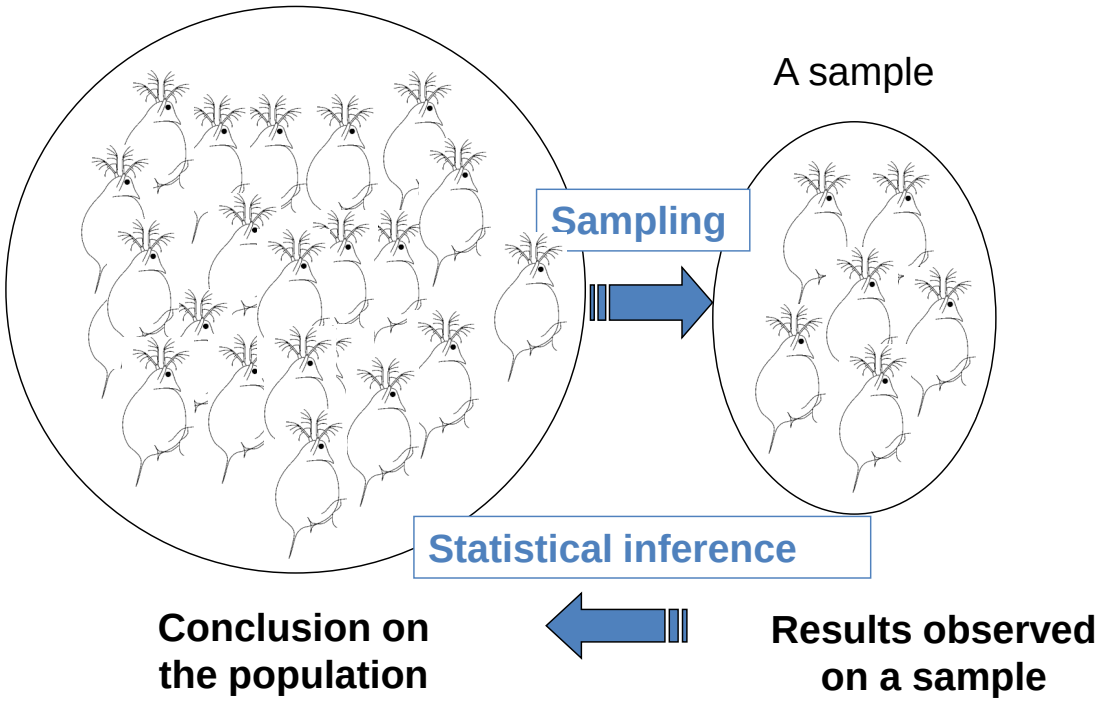
Search of the best fit

Different criteria, such as maximum likelihood ($max_\theta(P(Y \mid \theta))$), may be used to choose the best fit values for the model parameters.



Generalization to population

Inference also implies generalization of a result from a sample to population, and the calculation of uncertainty in the estimated parameters, especially uncertainty due to sampling error.



A very simple example

Estimation of a survival rate (probability to survive) for studied organisms in fixed conditions
(data : $y = 24$ survivals among $n = 100$ organisms)
Model :

- no deterministic part
- no covariate
- stochastic part : $y \sim binomial(p, n)$

This model is characterized by only one parameter (p).
In the following, for our purpose to remain general, we will continue to name θ the vector of model parameters.

Point estimate of θ : $\widehat{\theta}$

frequentist framework

θ is assumed fixed but unknown

It is estimated by one of the following methods :

- moment matching
- maximum likelihood ($max_\theta(P(Y \mid \theta))$)
- sum of squared deviations minimization

(Different methods may lead to the same estimation in some cases).

In our example, the estimated survival rate is $\frac{24}{100} = 0.24$

Interval estimate of θ

The calculation of a confidence interval (generally a 95% interval)
is based on imagining repeated sampling from the model :

Definition of a confidence interval

If we repeatedly obtained samples of size n from the population and constructed a 95% confidence interval for each, we could expect 95% of the intervals to contain the true value of the parameter.

In average, when we calculate 95% confidence intervals, 1 out of 20 does not contain the true value of the parameter we want to estimate.
In our example, the 95% confidence interval of the survival rate is [0.16;0.34]

Hypothesis test concerning θ

Ex. : comparaison of the survival rate in a contaminated medium to the one observed in the control $\theta = \theta_0$?
Calculation, under H_0 $(\theta - \theta_0 = 0)$,
of the p-value

$$\text{p-value} = P(|\widehat{\theta - \theta_0}| > |\widehat{\theta - \theta_0}|_{obs} | H_0)$$

p-value definition

Assuming the null hypothesis true, probability (in the frequentist meaning, imagining the repeated sampling) to obtain an estimated difference greater that the one observed on the sample.

if p-value $< \alpha$, H_0 is rejected
(generally $\alpha = 5\%$)
and the difference is said to be significant.

Common abusive interpretation of a p-value

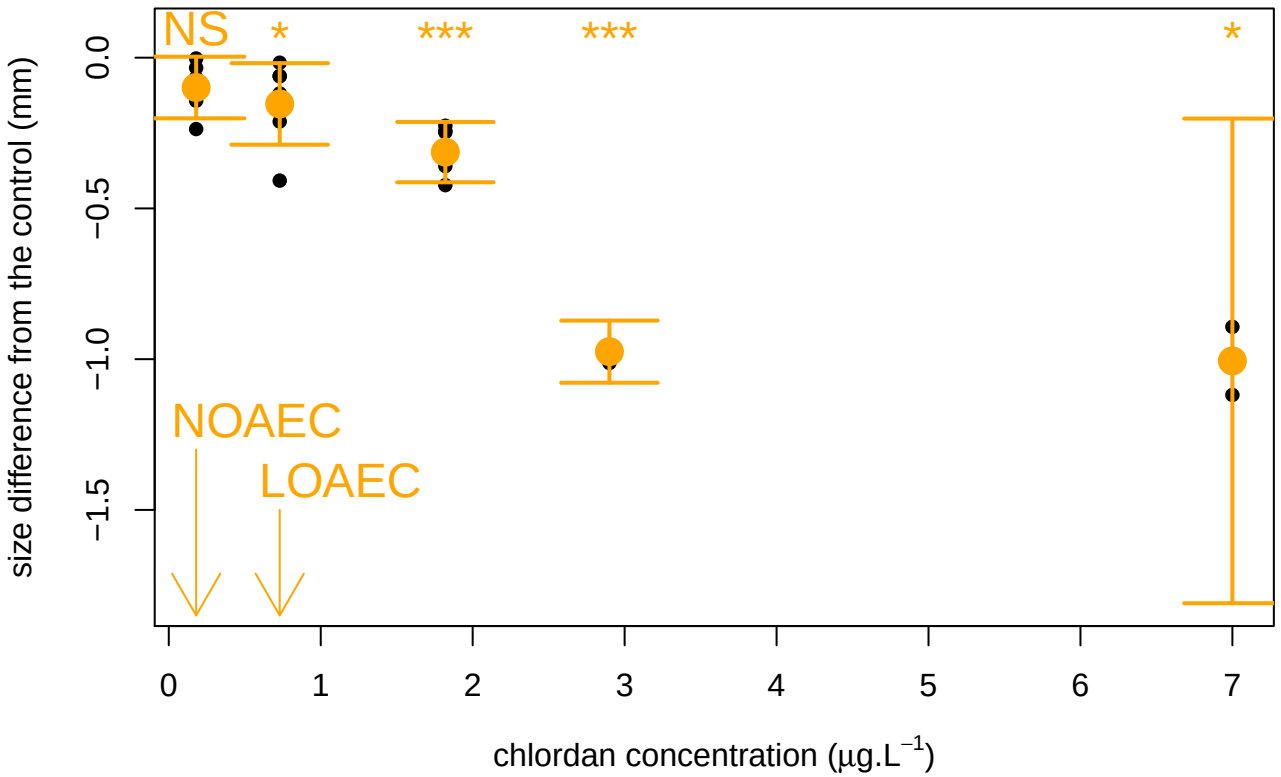
We should not accept the null hypothesis ($H_0 : \theta - \theta_0 = 0$) when p-value $> \alpha$ without taking into account type II error (β risk).
A difference may be non significant due to lack of power of the test,
and power/sample size calculation requires the prior definition of the minimal difference you want to be able to detect (not often done).

SO BE CAREFUL!

An hypothesis test does generally not allow us to accept H_0.

Abusive interpretation of p-values while defining NOECs

This misinterpretation of p-values causes many problems in the use of the NOEC as a toxicity threshold (*e.g.* the less data are available, the higher the estimated value of the threshold)



Bayesian estimation of θ

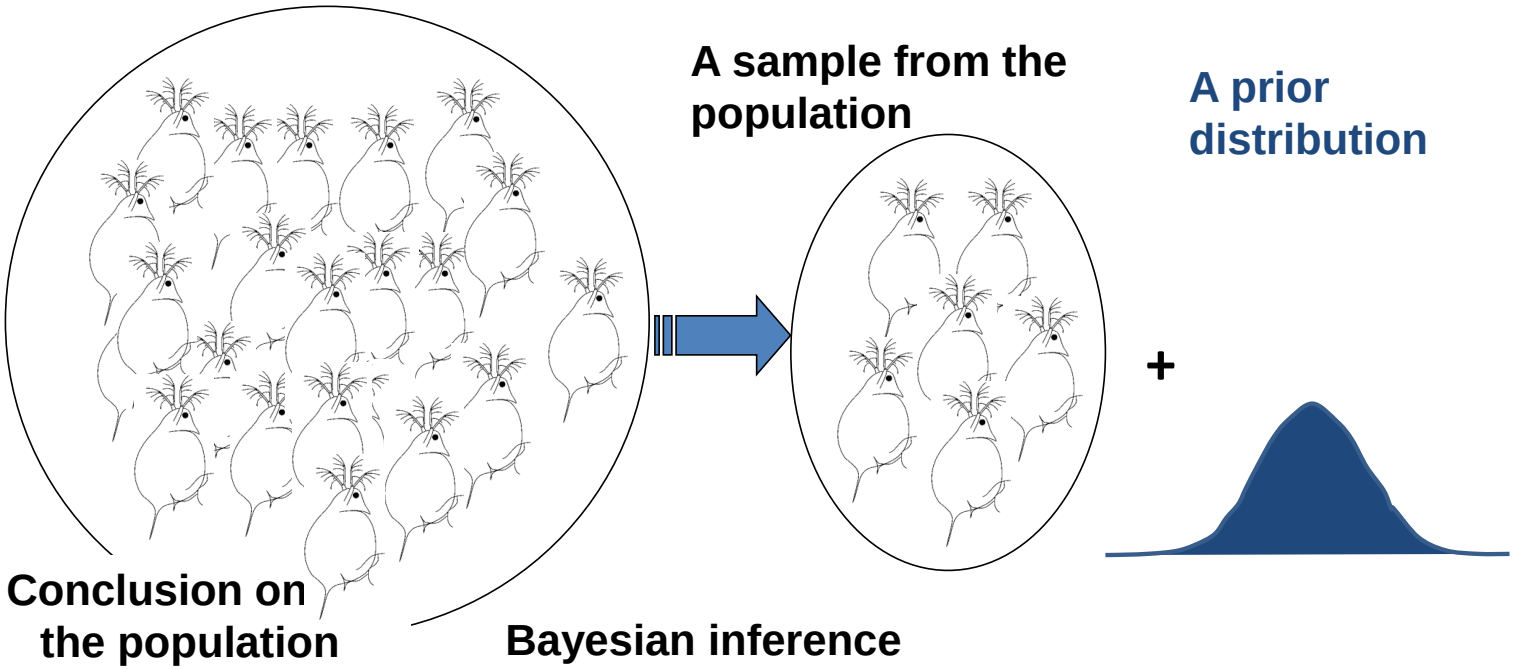
Bayesian framework

θ is supposed uncertain, and its uncertainty characterized by a probability distribution (subjective meaning of a probability, degree of belief)

- Prior distribution : $P(\theta)$ more or less informative
- Posterior distribution : $P(\theta|Y)$
  calculated using Bayes theorem :
  from the prior distribution and the likelihood function $P(Y|\theta)$

$$P(\theta|Y) = \frac{P(Y|\theta) \times P(\theta)}{P(Y)} \propto P(Y|\theta) \times P(\theta)$$

Bayesian inference



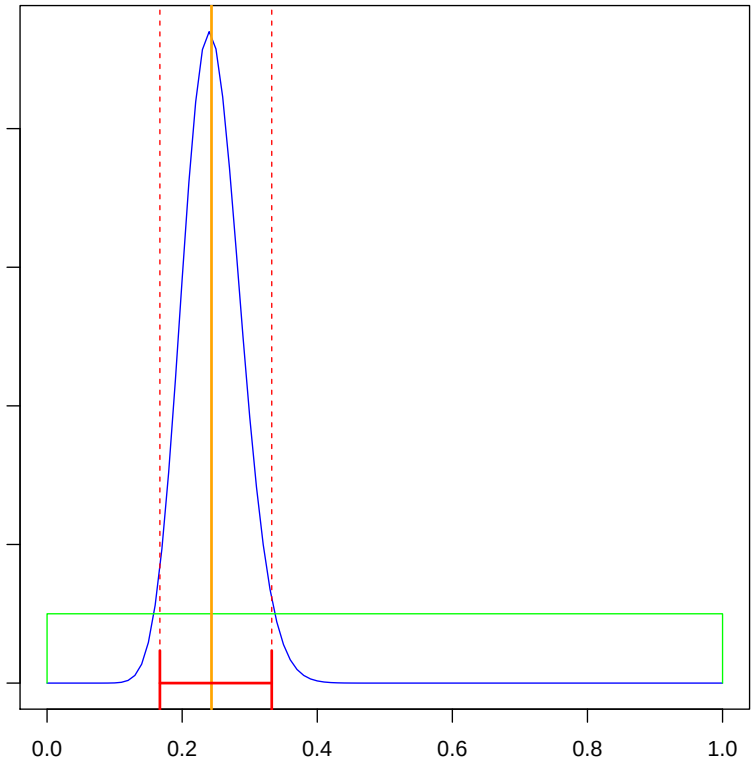
The posterior distribution = the conditional distribution of the unknown parameters given the observed data

Use of the posterior distribution for parameter estimation

- **Point estimate :**
  mean, median or mode of the posterior distribution
- **Interval estimate :**
  definition of a **credible interval** (or bayesian confidence interval)
  from posterior distribution quantiles (2.5% and 97.5% quantiles for a 95% credible interval).
  Such an interval is easier to interpret than a frequentist confidence interval : the probability that the parameter lies in a 95% credible interval is 95%.
- **Hypothesis test :**
  It is no more necessary to calculate any p-value : one can make decisions from posterior distributions.

Example : bayesian estimation of a survival rate

- likelihood family function: $binomial(p, n)$
- data : $y = 24$ survivals out of $n = 100$
- prior distribution : $uniform(0, 1) = beta(1, 1)$ non informative
- posterior distribution (analytically known in that simple case) : $beta(y + \alpha, n - y + \beta)$ here $beta(y + 1, n - y + 1)$ $= beta(25, 77)$



- point estimate : 0.24
- 95% credible interval : [0.17; 0.33]

♡♡♡ Remember ! ♡♡♡

- **Frequentist framework**
  - **θ is supposed fixed but unknown**
  - Parameter inference only uses observed data
  - Confidence intervals are defined imagining repeated sampling from the model, the probability being associated to the relative frequency of occurence of an outcome.
- **Bayesian framework**
  - **θ is considered as a random variable, associated to a probability distribution, in the subjective meaning of a probability (degree of belief)**
  - Parameter inference uses both observed data and a prior information (prior distribution)
  - Credible intervals are defined from the posterior distribution, and can be easily interpreted : 95% is the probability that a parameter lies in a 95% credible interval.

Analytical calculation of the posterior distribution

$$P(\theta|Y) = \frac{P(Y|\theta) \times P(\theta)}{P(Y)} \propto P(Y|\theta) \times P(\theta)$$

This calculation is often tricky.

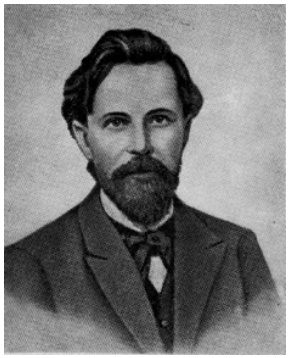
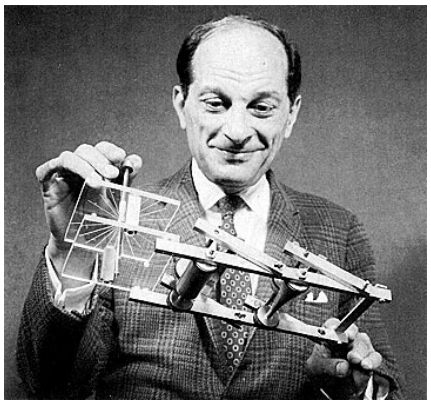
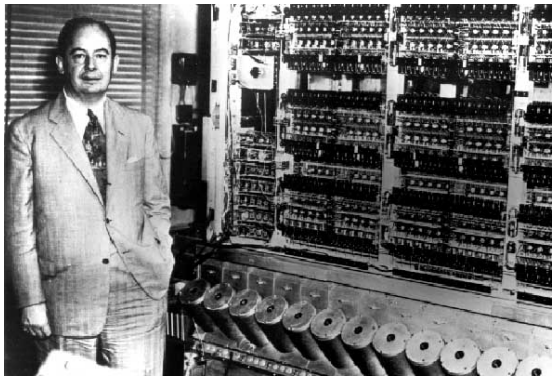
An analytical result exists only in some cases

$\Rightarrow$ strong limitation of the use of Bayesian framework.

For a long time, its use has been limited to simple cases.

Estimation of the posterior distribution by numerical simulations

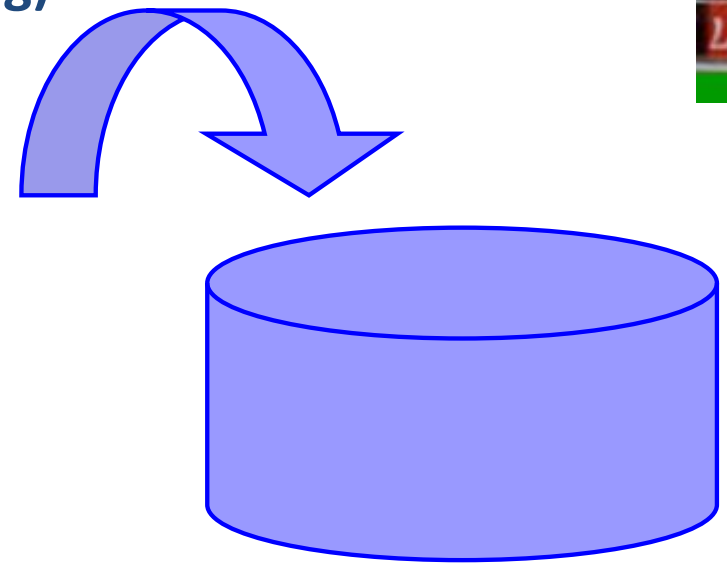
Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution (in Bayesian inference, the posterior distribution).

Markov Chain (A. Markov) Monte Carlo (S.Ulam and J. Von Neumann)

MCMC algorithms

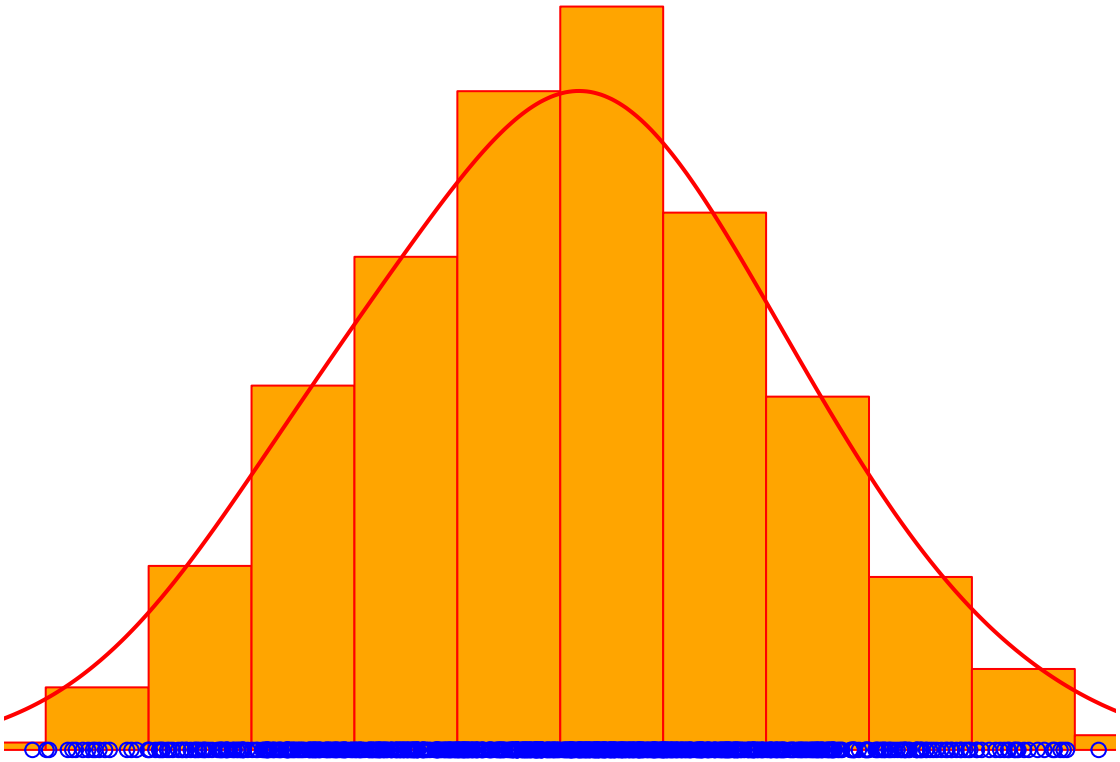
Objective of MCMC algorithms in Bayesian inference :
to draw a random sample that converges to the posterior distribution (in a distribution meaning)



MCMC simulations

MCMC simulations

Posterior distribution characterization from MCMC simulations

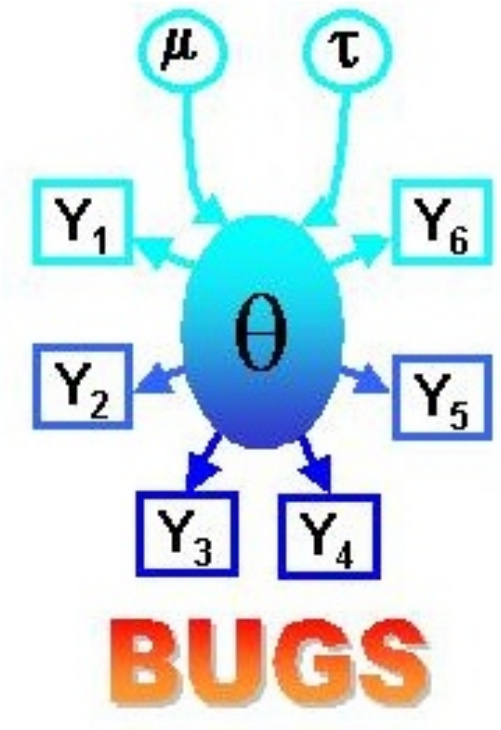


Algorithms and software

- **Algorithms**
  - Metropolis - Hasting algorithm published in 1953 by N. Metropolis and generalized in 1970 by W.K. Hastings
  - Gibbs sampling algorithm published in 1984 par S. Geman et D. Geman (special case of Metropolis - Hasting algorithm, easier to implement)
- **BUGS project** (since 1989)
  Bayesian inference **U**sing **G**ibbs **S**ampling
  Flexible tools facilitating the implementation of Bayesian inference with any user-supplied models, using MCMC algorithms
  Main tools :
  - Winbugs
  - Openbugs
  - JAGS

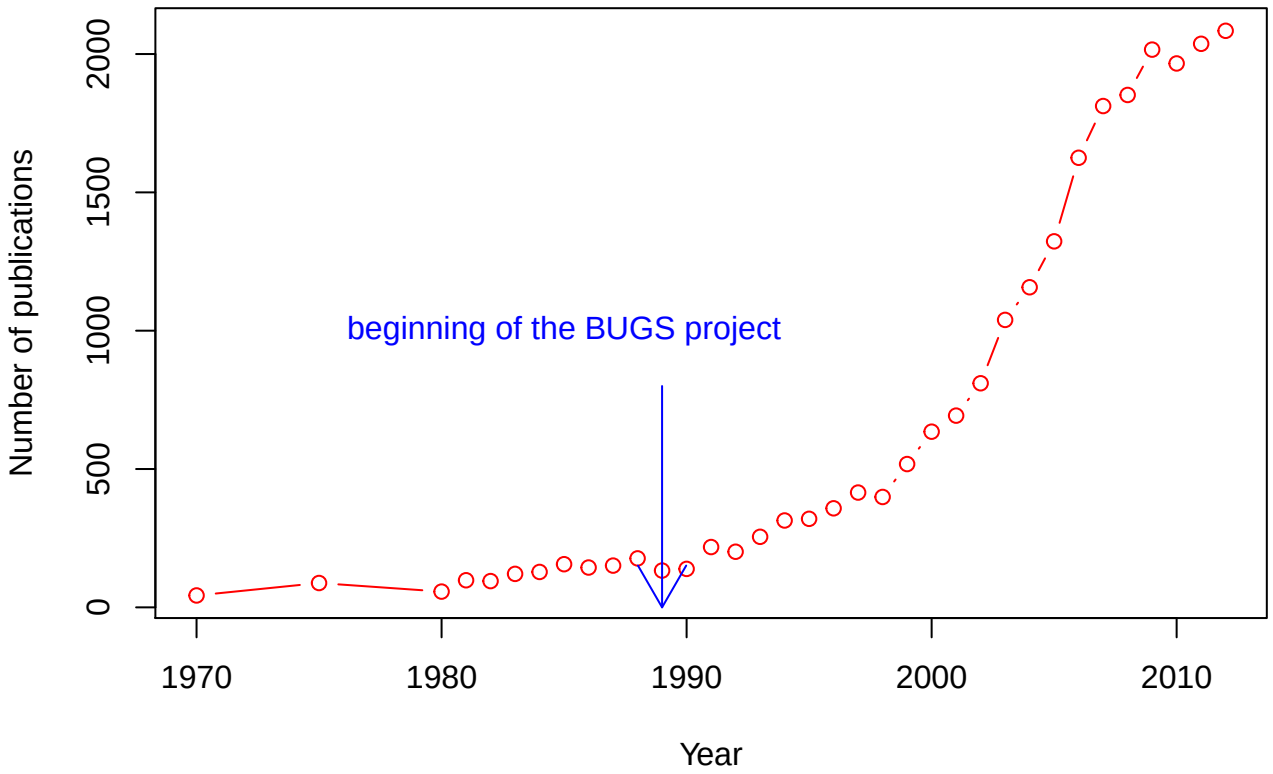
Central website of the BUGS project

http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml



Use of Bayesian inference in last decades

Search of papers containing the word **Bayesian** in their title (ISI Web of knowledge)



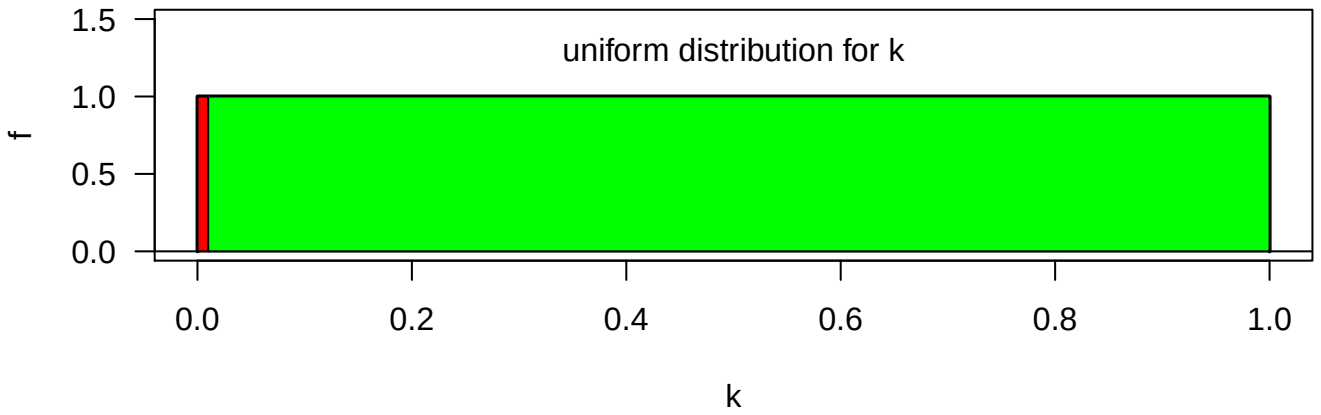
How to choose priors ?

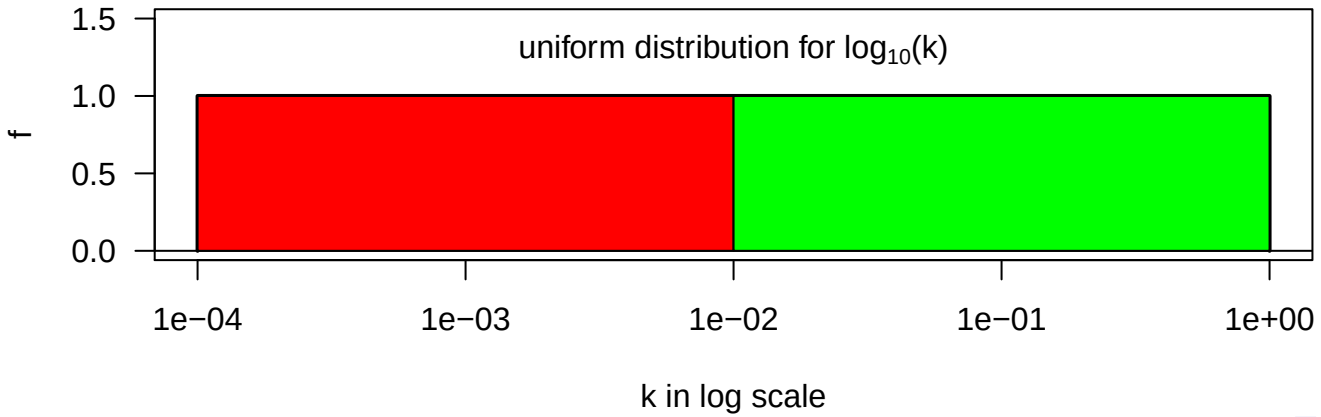
A prior may be more or less informative depending on what you know **before looking at the data**
Two main questions :

- **Range of possible values for each model parameter ?**
- **Shape of the distribution of each parameter on its range ?**
  Ex. of a parameter k varying on the range $[10^{-4}, 1]$
  - $k \sim Unif(10^{-4}, 1)$ implies that
    $Pr(k \in [10^{-2}, 1]) = 100\times\ Pr(k \in [10^{-4}, 10^{-2}])$
  - $log_{10}(k) \sim Unif(-4, 0)$ implies that
    $Pr(k \in [10^{-2}, 1]) = Pr(k \in [10^{-4}, 10^{-2}])$
    In such a case, this second choice is generally more acceptable

Impact of the shape of priors : illustration

Visualization of probabilities :
$Pr(k \in [10^{-4}, 10^{-2}])$ and $Pr(k \in [10^{-2}, 1])$





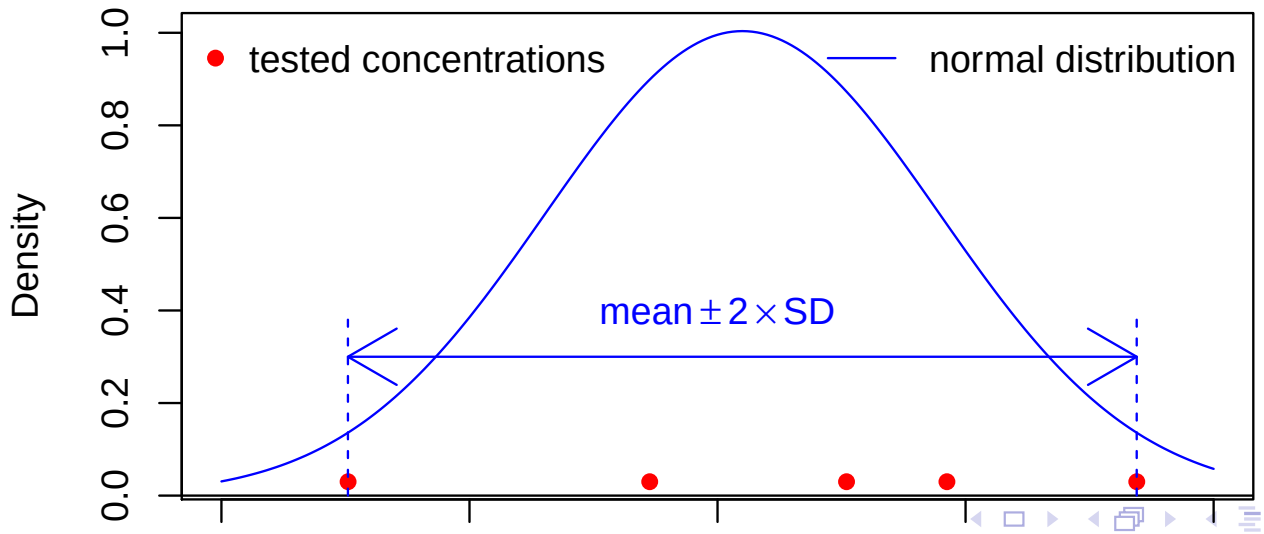
Some advices to choose priors

- **for a non-informative or vaguely informative distribution**
  a large uniform distribution can generally be used
  - directly on the parameter if its order of magnitude is known
  - on the log-transformed parameter if its order of magnitude is unknown
- **for an informative distribution**
  a distribution with an infinite support is preferable, as the posterior distribution is constrained by the prior support (e.g. a normal distribution on the parameter, log-transformed or not, truncated if necessary)

The prior is what you know before making the experiment.
So never use observed data in order to define priors. But you may use the experimental design (e.g. values of tested concentrations).

Example of definition of priors from tested concentration

Set of tested concentrations: 0.18, 0.73, 1.82, 2.9 and 7.
As the range between the minimum (C_{min}) and the maximum (C_{max}) is large, we may log_{10}-transform these values:
-0.745 -0.137 0.260 0.462 0.845
and assume a normal distribution for $log_{10}(EC50)$ centered on $\frac{C_{min}+C_{max}}{2}$ with a probability of 95% to lie between (C_{min}) and (C_{max}).



Posterior check of priors

- **comparison priors/posteriors**
  numerical and/or graphical comparison of priors and posteriors in order to check that the priors are not too informative, (do not constraint too much the posteriors).
- **sensitivity analysis to prior choice**
  repeating the inference by modifying priors in order to check the robustness of results to choices concerning priors

Comparison of priors and posteriors (4p-loglogistic model with *Daphnia magna* - chlordan growth data)

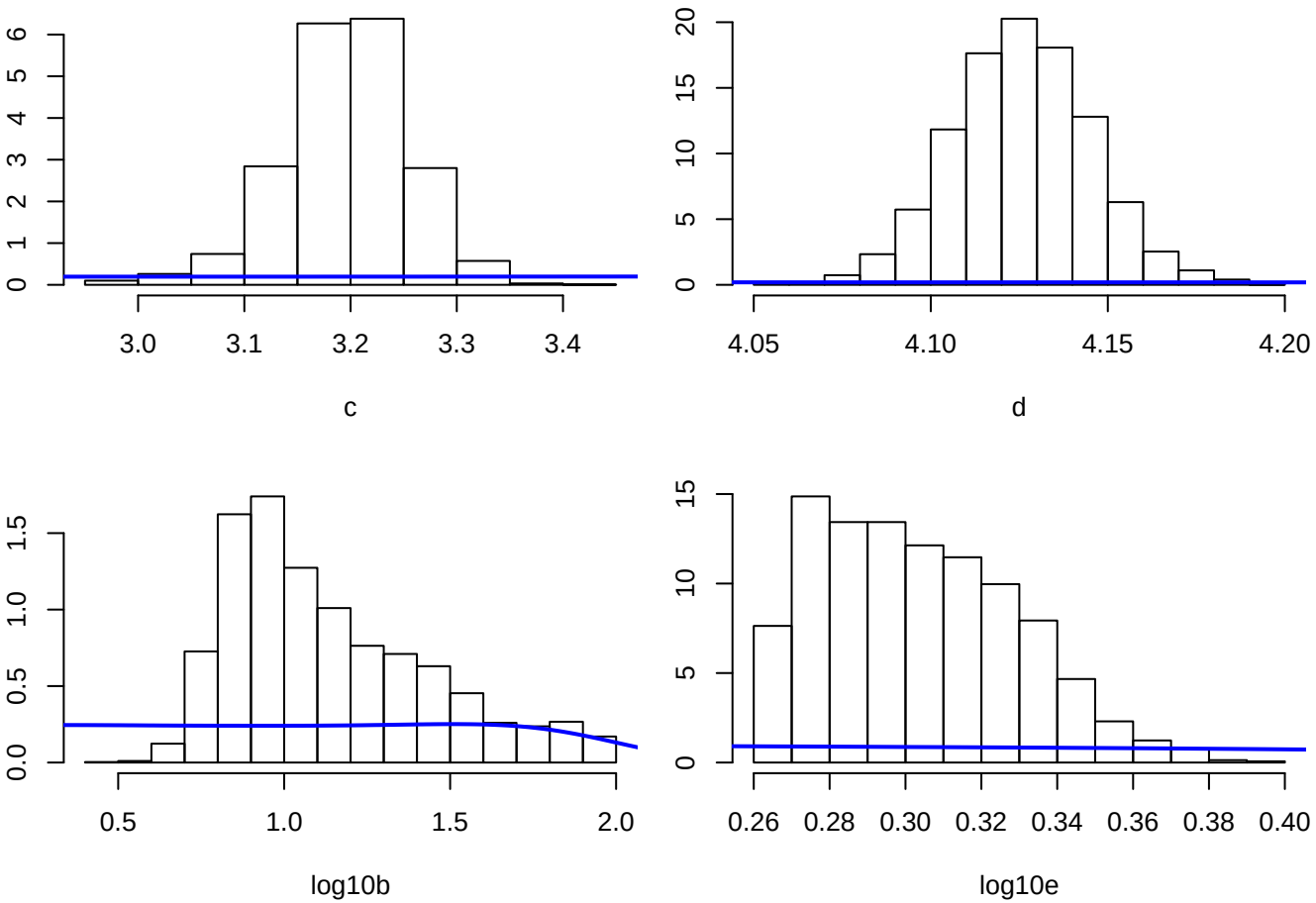
- **Marginal prior distributions of parameters**

	2.5%	25%	50%	75%	97.5%
c	0.1444	1.237	2.4809	3.754	4.861
d	0.1224	1.213	2.4629	3.662	4.872
log10b	-1.9059	-1.013	-0.0189	1.004	1.903
log10e	-0.7450	-0.193	0.0682	0.336	0.826
sigma	0.0482	0.502	0.9856	1.503	1.959

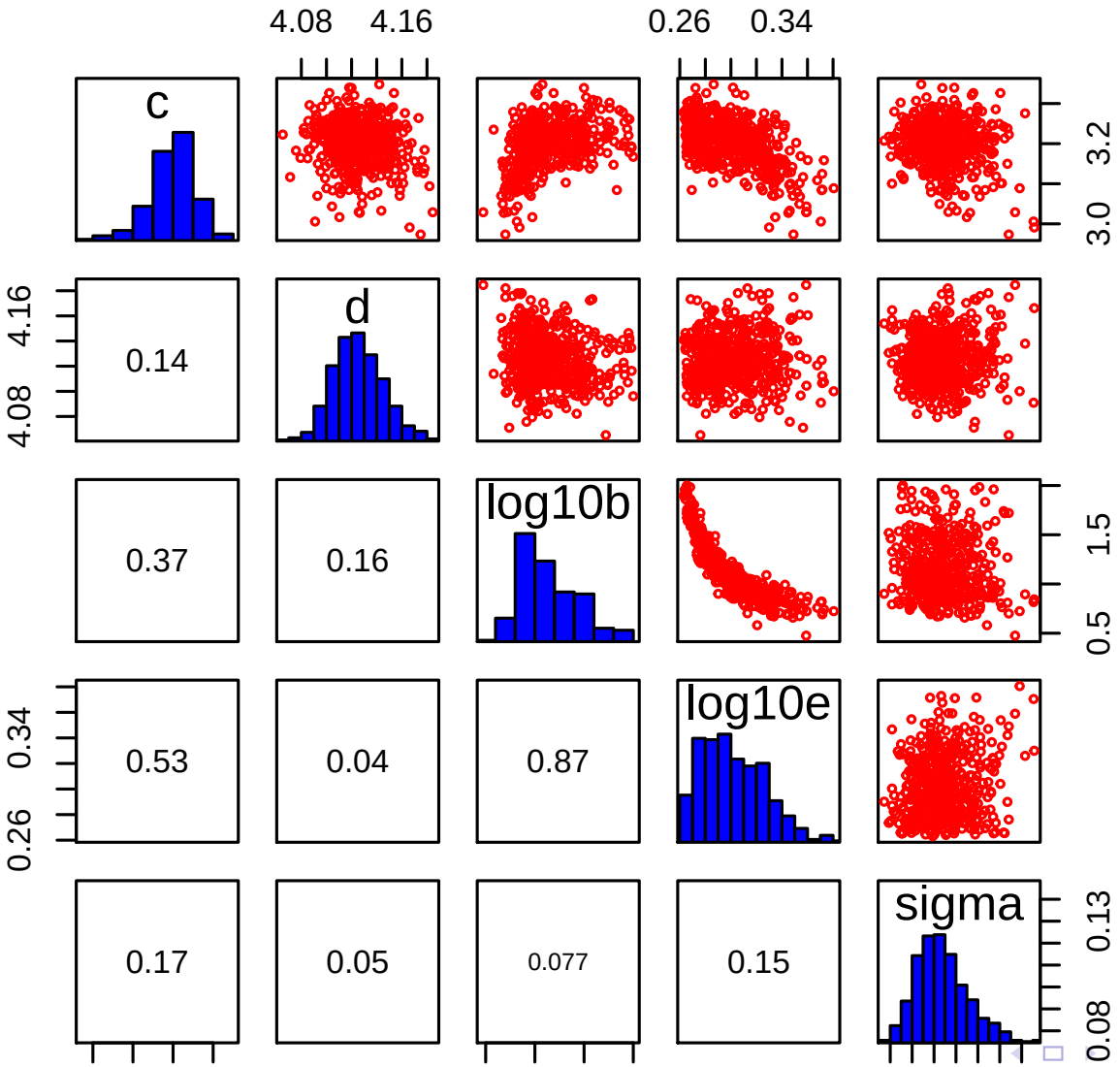
- **Marginal posterior distributions of parameters**

	2.5%	25%	50%	75%	97.5%
c	3.0702	3.1611	3.199	3.237	3.306
d	4.0873	4.1129	4.126	4.139	4.165
log10b	0.7254	0.9005	1.058	1.328	1.869
log10e	0.2661	0.2816	0.301	0.322	0.358
sigma	0.0833	0.0951	0.103	0.112	0.132

Visual comparison of priors and posteriors

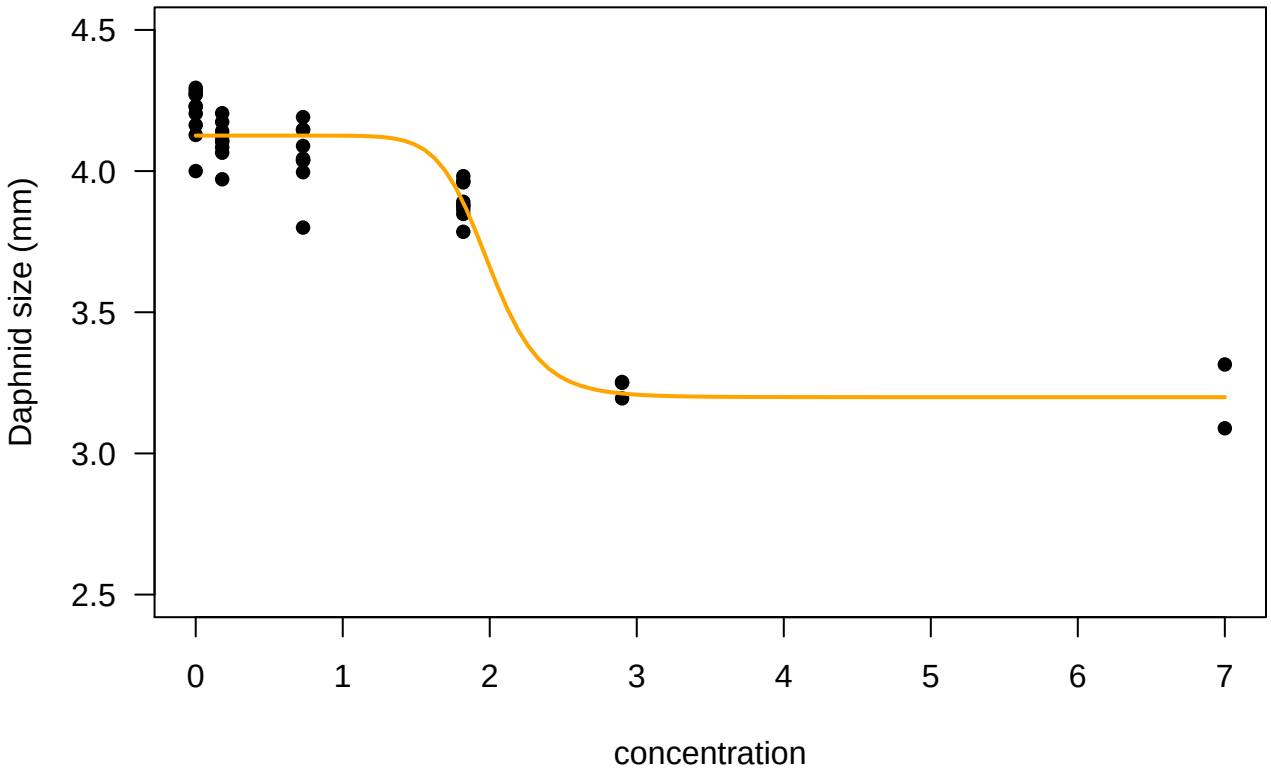


Joint posterior distribution of parameters



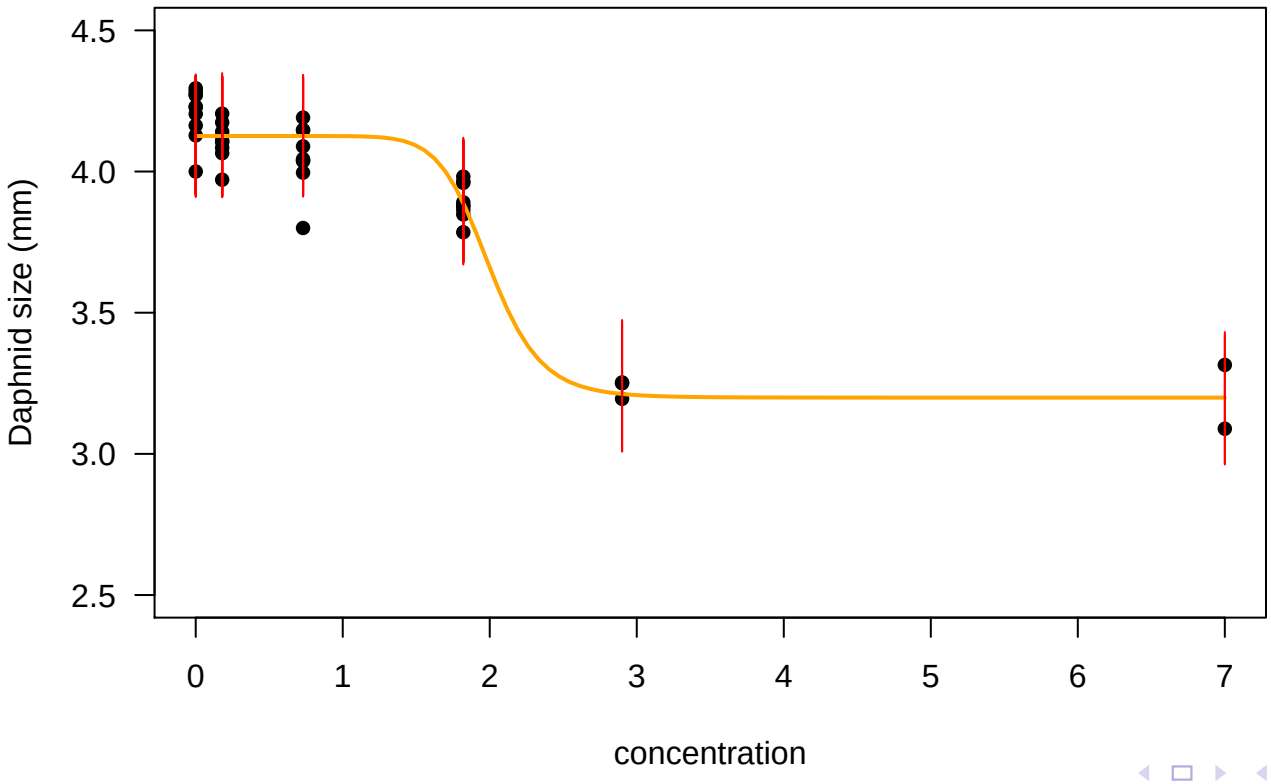
Prediction using point estimates of parameters

predicted growth curve (using median of posteriors)



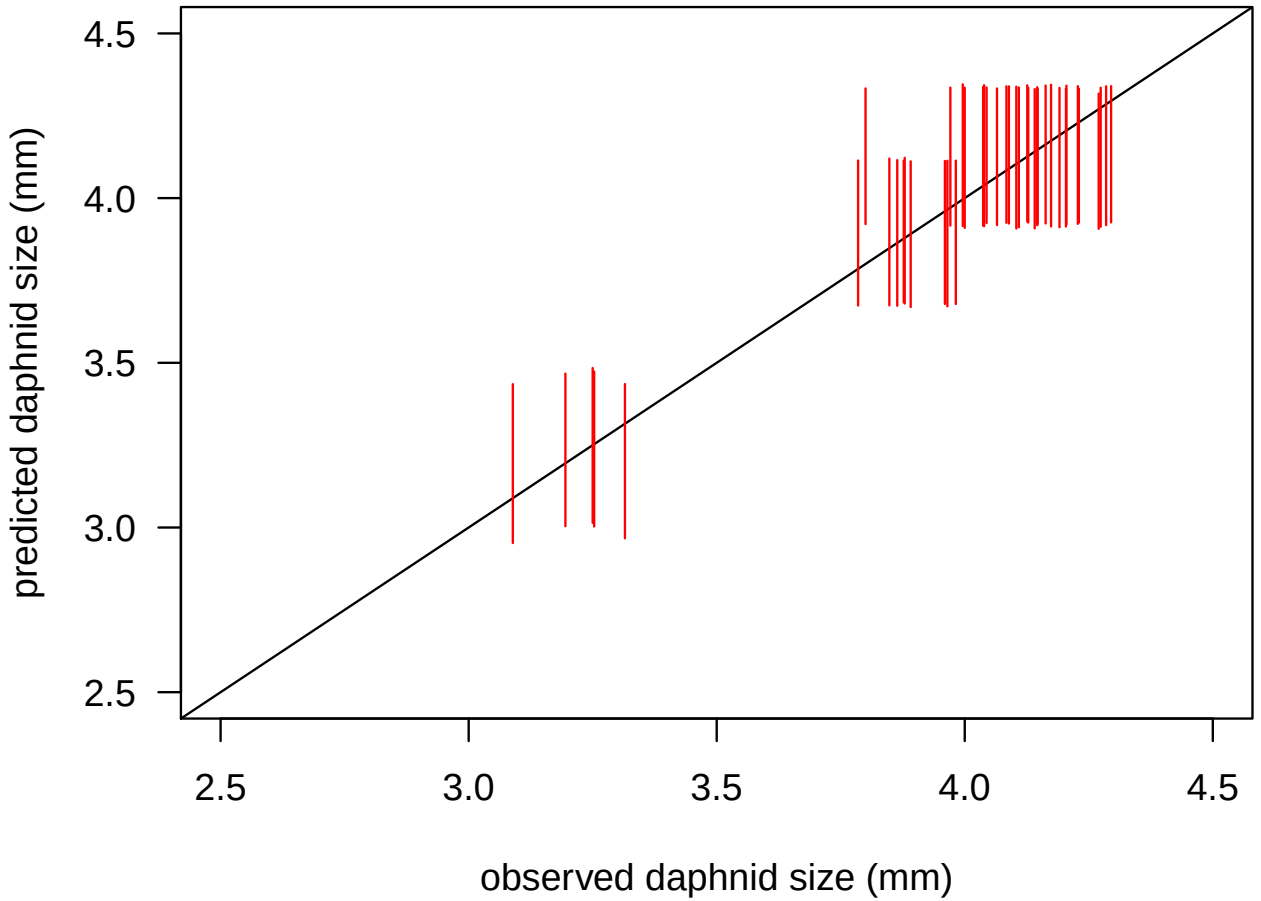
Predictions with credibility intervals (by simulations in the joint posterior distribution)

predicted growth curve with 95% crediblity intervals for each concentration compared to **observed values**



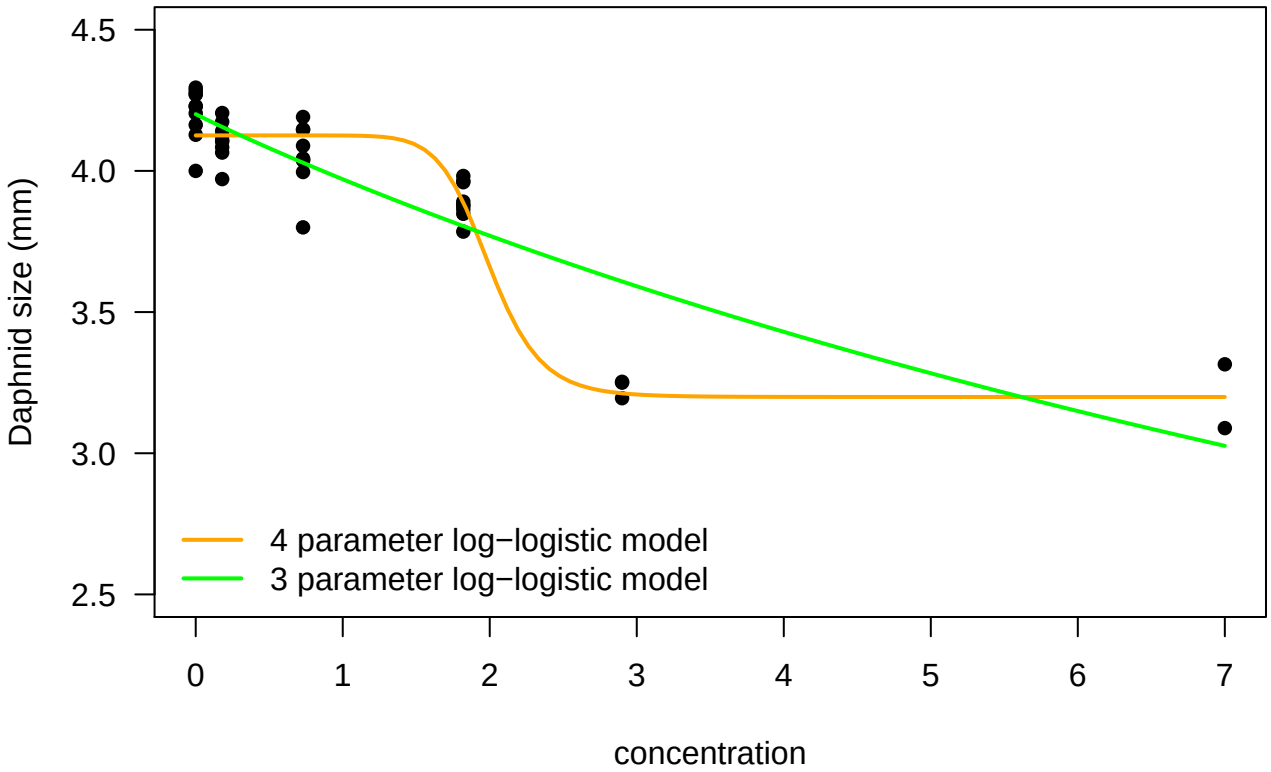
Another representation for predictive checking

95% crediblity intervals (CIs) against **observed values**:
95% CIs are expected to contain 95% of the observations.



Comparison of models

Comparison of two models of different complexities: example of log-logistic models with 4 or 3 parameters



DIC : Deviance Information Criterion

Information criteria commonly used to compare fits of two models:
Deviance penalized by model complexity,
Deviance : $D(Y,\theta) = -2 * log(P(Y|\theta))$
DIC : $DIC = D(Y,\overline{\theta}) + 2 * p_D$ with $p_D = \overline{D(Y,\theta)} - D(Y,\overline{\theta})$
Generalization of the Akaike criterion ($AIC = D(Y,\widehat{\theta}) + 2 * n_{par}$),
especially suited to compare hierarchical models.
The model with the smaller DIC is prefered.
Calculation of DIC3p - DIC4p with `rjags` on the example

```
> diffdic(DIC3p,DIC4p)

Difference: 31.6
Sample standard error: 14.3
```

The model with 4 parameters would be prefered.

When to use Bayesian inference ?

Why make simple when we can complexify ?

Incorporation of prior information

A prior information is often available, which may be very useful, especially when data are sparse.

- prior biological knowledge of parameters of biologically based models
- prior knowledge from previous experiments

When experimental data are insufficient to estimate all the model parameters, it seems better to define a prior for each parameter (according to its prior knowledge) that to arbitrarily fix some of them.

Fitting otherwise intractable models

Inference with complex models is often easier within a Bayesian framework

We often use Bayesian inference just to meet our needs, not out of pure ideology

Bayesian inference is of special interest for

- models with **latent variables**
- **hierarchical** models
- **non gaussian error** models
- **non-linear** models

An example in ecotoxicology

Fit of a non-linear model to survival data, the same organisms being followed over time.
The deterministic part of the model links the survival rate to the time.

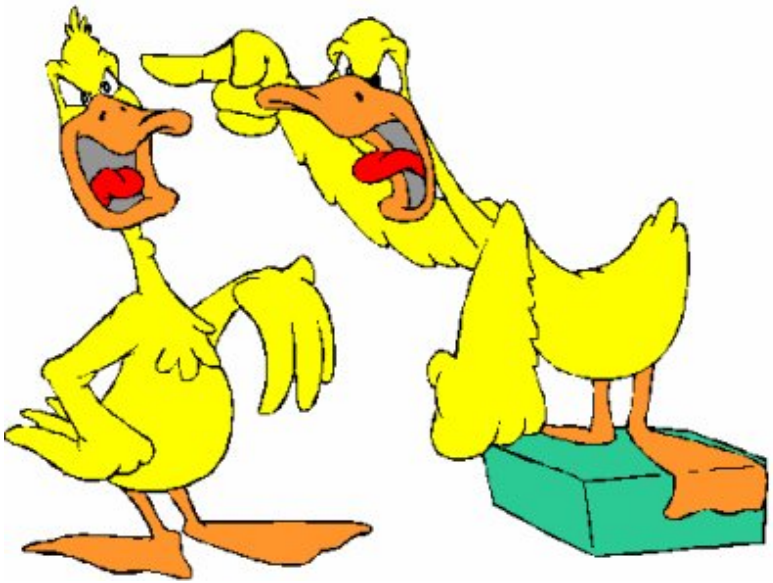
Some difficulties

- non-linearity of the model
- non gaussian error model (quantal variable - binomial distribution)
- dependence of the successive measurements (conditional binomial model or equivalent multinomial model)

It is much simpler to implement such a model using one of the BUGS software than to implement it by maximum likelihood.

The end of an old quarrel ?

Is the old quarrel that divided frequentist and Bayesian statisticians over ?



Conclusion by Bradley Efron

19th Century science was broadly Bayesian in its statistical methodology, while frequentism dominated 20th Century scientific practice. This brings up a pointed question: which philosophy will predominate in the 21st Century? One thing is already clear: statistical inference will pay an increased role in scientific progress as scientists attack bigger, messier problems in biology, medicine, neuroscience, the environment, and other fields that have resisted traditional deterministic analyses. A combination of frequentist and Bayesian thinking will be needed to deal with the massive data sets scientists are now bringing us.