



# Plan

- **1 HOUR:**

- 10 min welcome & klaxoon link <https://app.klaxoon.com/join/SKRBXHD>
- 30 min intro
- 10 min intro to software tools
- 5 min Q&A
- 5 min Colab tutorial

- **2 HOURS:**

- 20 min growth+censoring examples colab /Jukka
- 30 + 30 min inactivation+growth examples /Kento & Hiroki
- 30 min dose-response examples /Anne
- 10 min Q&A Optional examples.

← Give comments here!



# What is your interest?

- Do you have specific problems for which Bayesian methods could apply?
- What would you like to know?
- Have you used Bayesian inference?



Give comments: <https://app.klaxoon.com/join/SKRBXHD>





**RUOKAVIRASTO**  
Livsmedelsverket • Finnish Food Authority

# Bayesian Modelling intro, ICPMF12 workshop

[jukka.ranta@foodauthority.fi](mailto:jukka.ranta@foodauthority.fi)

PhD, res.prof, docent (biometry)

Risk assessment unit

**ICPMF12**

13.6.2023

# Bayesian advantages in general



- Pros:
  - Extremely **flexible** for complex estimation problems.
    - E.g. nested structures, missing data, unobservable variables and measurement errors.
    - Multiparameter inference, combining multiple/all data.
  - **Unified** 'Single step' approach for estimation.
    - Not estimating each parameter separately with various methods – but all at once.  
→ quantifies the combined uncertainty.
  - Using probability to directly address the uncertain quantity, conditionally, given the known data.
    - **Updating probabilities** with new data.
    - Quantification of both **variability and uncertainty**: either integrated, or separated 2D simulation.
    - **Explicit assumptions** by conditional probabilities fit for modular problem formulation.



# Bayesian challenges commonly

- Cons:
  - **Computationally** demanding.
    - Need special software, practise, some knowledge of probability calculus/distributions.
    - Assessment of convergence of MCMC simulation.
    - MCMC simulation can be slow, (choice of initial values).
    - Numerical problems could occur within simulations.
  - Maybe too easy to try too large models with too many parameters?
  - Model **selection**, which to choose? (DIC? AIC? WAIC? Posterior predictive?)
    - Some models are easier for MCMC sampling than others.
    - More parameters or less parameters?
    - Choice of prior distribution. Uninformative/informative priors?
    - Cross-validation for model choice – even more computations.



# Bayesian statistics: focus on *distributions*, not point estimates.



- Provides the probability  $p(\Theta | \text{data})$  for the quantity  $\Theta$  we want to estimate, but often by using approximate computation (MCMC, Monte Carlo).
  - Classical frequentist statistics often gives exact  $p(T(X) | \Theta)$  for a *test statistics*  $T(X)$ , conditionally on null hypothesis  $\Theta \in H_0$ .

*“Far better an approximate answer to the right question, than an exact answer to the wrong question” — John Tukey*

- Flexibility comes with computational price! 
  - Needs careful judgment, no automatic recipe.

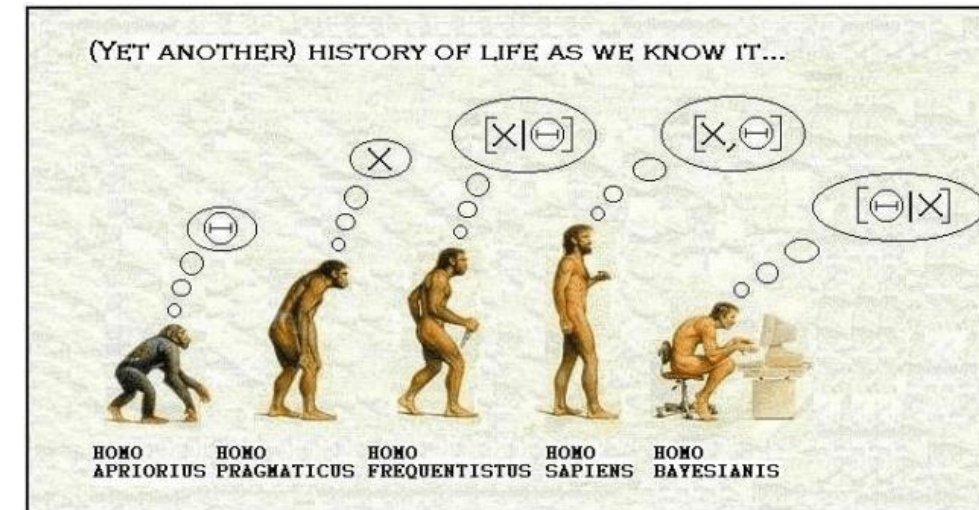
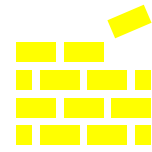
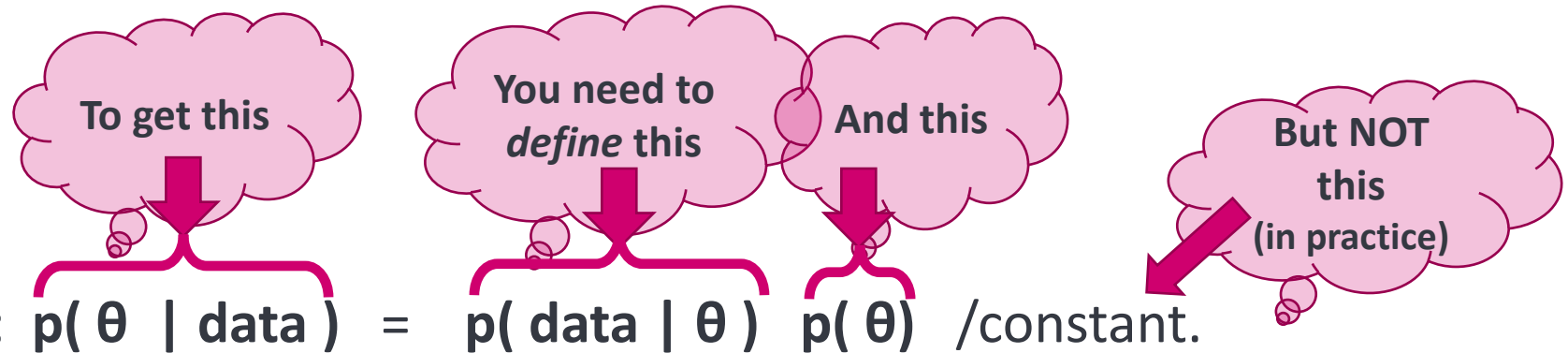


Image: Inference Bayes Frequentist (astro.it)

# Bayesian inference – essential building blocks



- Statistical inference based on probability theory:
  - **Target:** posterior distribution of the unknown parameter(s)
    - This is:  $p(\theta \mid \text{data})$  the conditional distribution of the parameter, given the data.
  - **Needs:**
    - $p(\text{data} \mid \theta)$  the probability model for all possible data, given the parameter.
    - $p(\theta)$  the prior distribution of the parameter.



- Bayes theorem:  $p(\theta \mid \text{data}) = p(\text{data} \mid \theta) p(\theta) / \text{constant}$ .

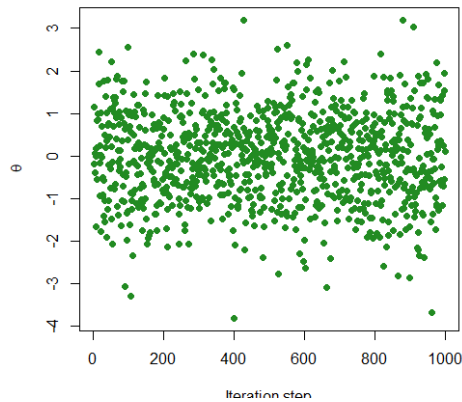


# MCMC – Markov chain Monte Carlo sampling

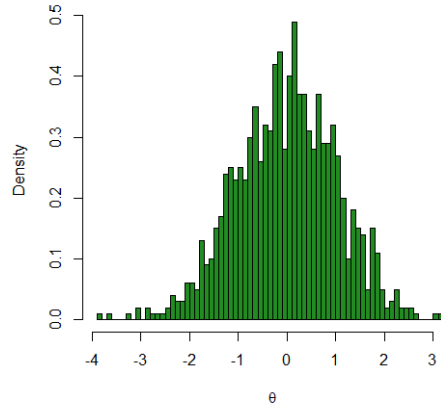
- Why? To get random samples representing the posterior distribution of  $\theta$

Independent  
Monte Carlo  
sampling

Monte Carlo sample from posterior distribution

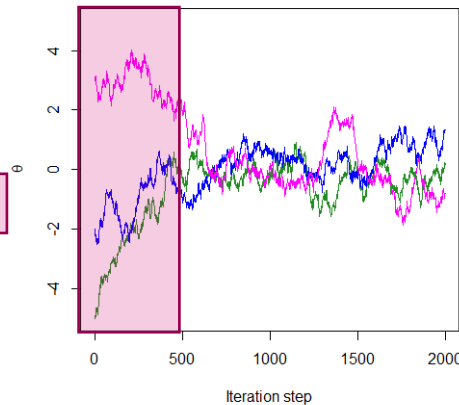


Sample from posterior distribution

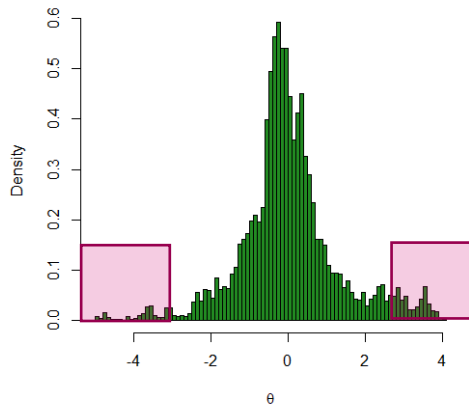


MCMC  
sampling  
with **burn-in**  
phase

MCMC sample from posterior distribution



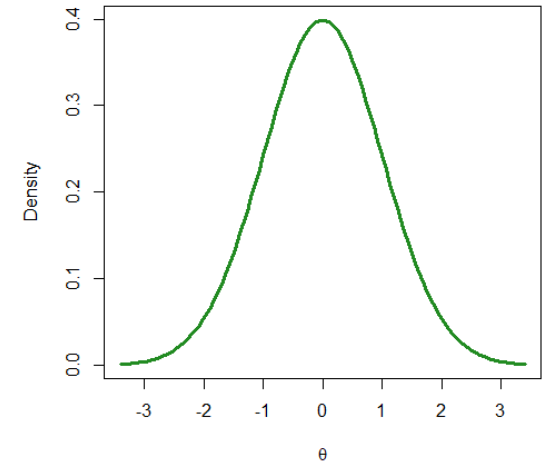
MCMC sample from posterior distribution



Converging ?

DIAGNOSTICS:  
multiple chains &  
Dispersed initial values.  
(BUGS/JAGS + tools)

**TARGET:**  
Posterior distribution







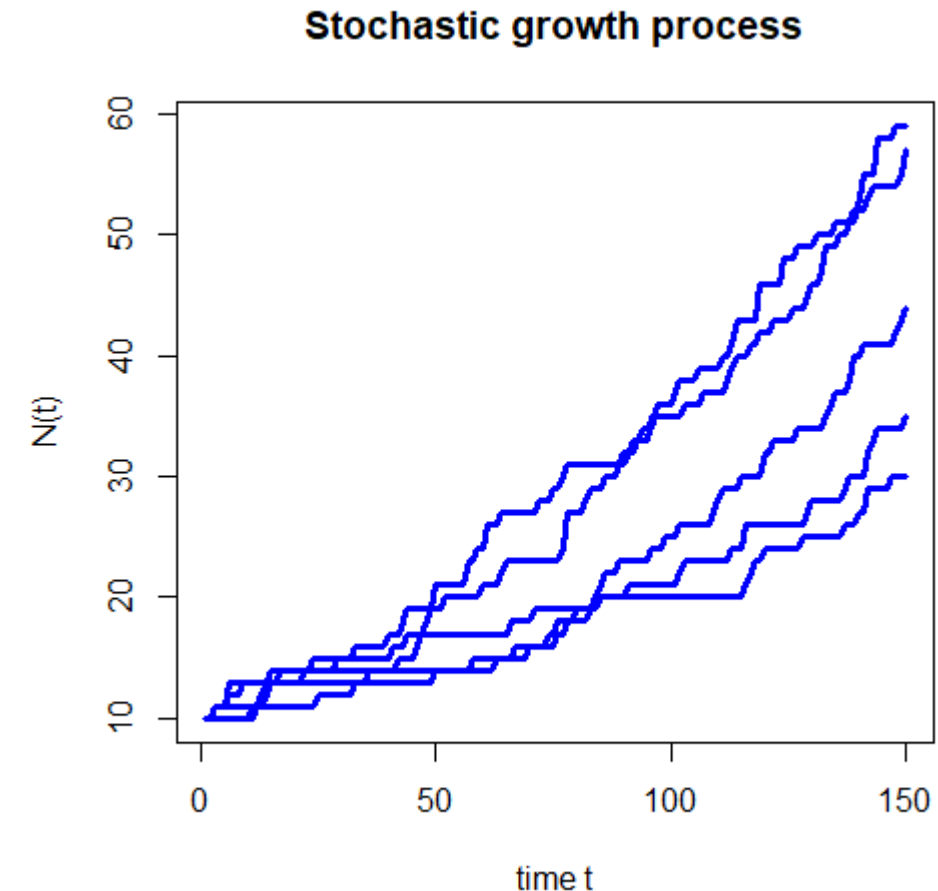
# Example: bacteria growth model

- Often deterministic functions (as 'primary models')
  - Exponential growth:  $N(t) = N(0)e^{\mu t} \Leftrightarrow \log N(t) = \log N(0) + \mu t$
  - Data for bacteria counts:  $N(t_1), N(t_2), \dots, N(t_n)$
  - Typical fitting: **search parameter value(s)** which minimize differences between data and model:  $N(t_i) - N(0)e^{\mu t_i}$  over all data points. Leads to a "best fit" model. (e.g. based on squared errors).  $\rightarrow$  only need a *minimizer* software.
  - How to quantify *uncertainty*?



# Stochastic growth model (as 'primary model')

- Bacteria counts  $N(t)$  (assuming no measurement errors)
  - Stochastic process model: e.g. 'pure birth model'
    - In continuous time.
    - Describes intrinsic stochastic variability over time.
    - Can only move upward!  $N(t_i) \leq N(t_j)$  when  $t_i \leq t_j$
    - Only one parameter in the model: growth rate  $\mu$ .
    - $N(t) - N(0) \sim \text{NegBin}(N(0), \exp(-\mu t))$  if  $N(0)$ ,  $\mu$ ,  $t$  are given.





# Bayesian inference for $\mu$ :

describe uncertainty of unknown parameter(s), based on data.

- With one observation  $N(50)=18$  from stochastic growth:

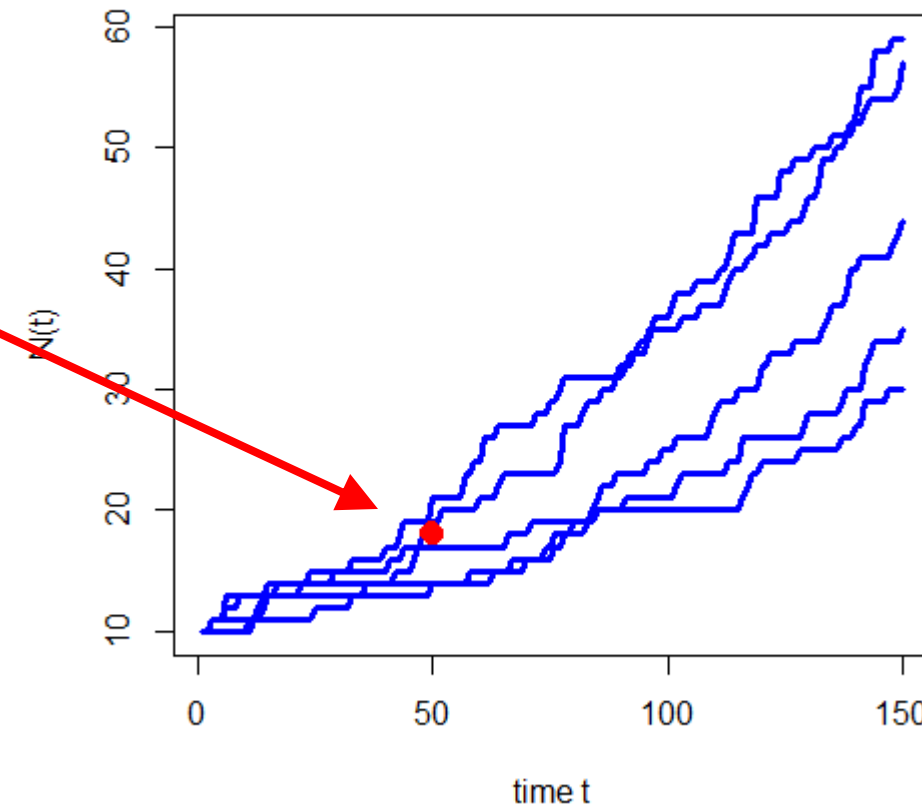
- Calculate the posterior distribution of growth rate  $\mu$ ?
- Assuming here:  $t = 50$ ,  $N(0)=10$ .

- $p(\mu \mid N(50)=18) = p(N(50)-N(0)=8 \mid \mu) p(\mu) / \text{constant}$ ,

- $p(N(50)-N(0)=8 \mid \mu) = \text{NegBin}(X=8 \mid 10, \exp(-\mu \times 50))$
- $p(\mu)$  = prior distribution, e.g.  $\text{Uniform}(0,1)$ .

{ Simple point estimate:  $N(0)\exp(\mu t) = E(N(t)) := 18$   
from which:  $\mu^* = \log(18/10)/50 = 0.01176$  }

Stochastic growth process





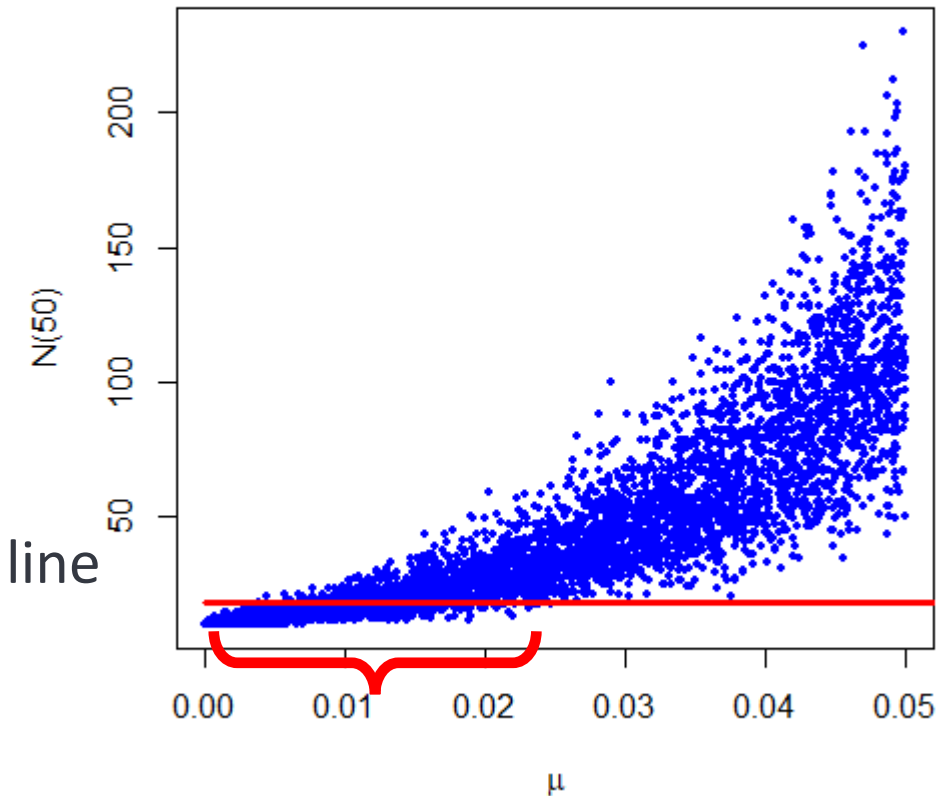
# Bayesian computations: simulation

- Bayesian model is a joint distribution:

$p(N(50), \mu)$  can be constructed as:  $p(N(50) | \mu)p(\mu)$

- From this it follows that we get  $p(\mu | N(50)=18)$  by **conditioning** on the value of variable  $N(50)=18$  (and assuming  $N(0)=10$ ).
- Hence, the posterior distribution of  $\mu$  is along the red line in the figure →
- Usually, obtained by using JAGS/BUGS/Stan.

Joint distribution  
 $p(N(50), \mu)$





# Other possible questions for estimation:

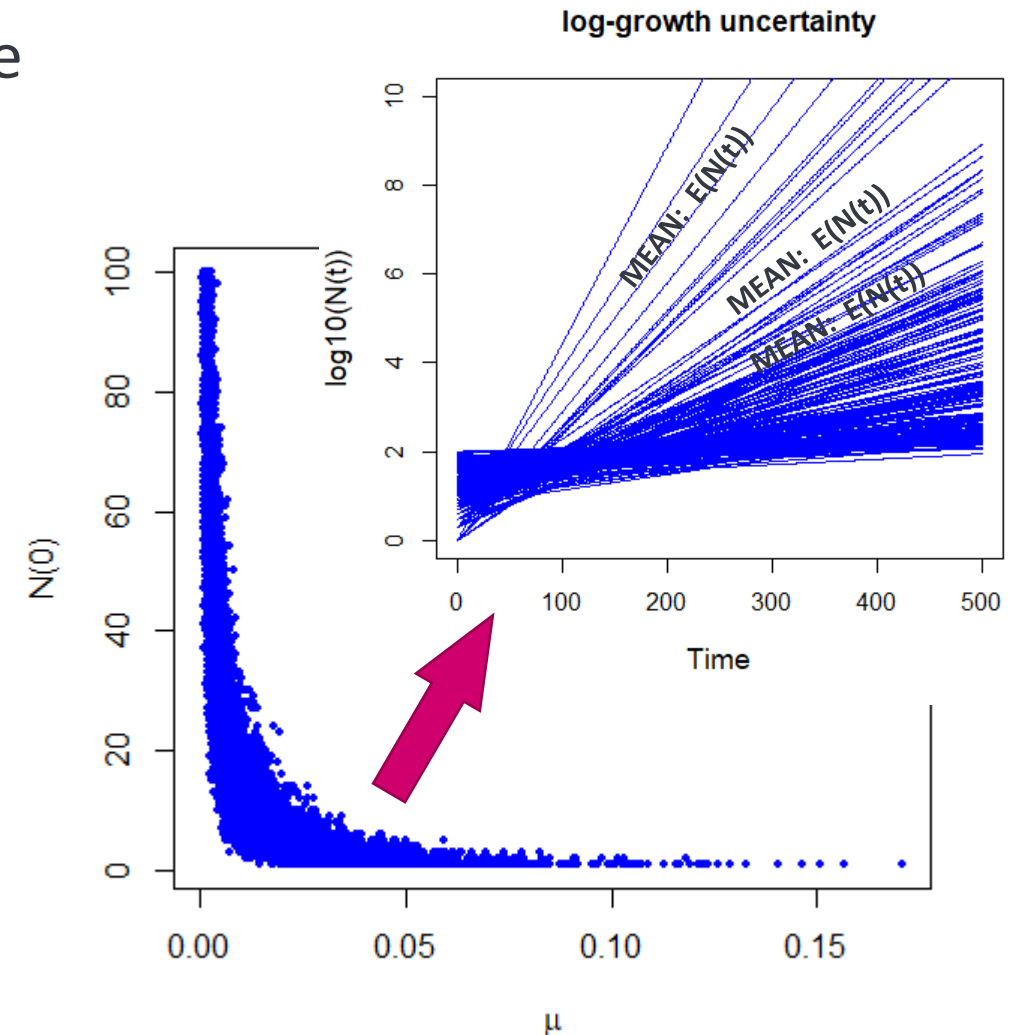
- Do you wish to estimate
  - the actual count  $N(t)$  ?
    - Compute:  $p(N(t) \mid N(0))$  , includes uncertainty and variability
  - the mean  $E(N(t)) = f(t, \mu)$  ?
    - Compute:  $p(\mu \mid N(0), N(t))$  , includes uncertainty
- **Initial count  $N(0)$ ?**, given assumption  $\mu = 0.01$ , and  $N(50)=18$ .  
Compute:  $p(N(0) \mid N(50)=18, \mu=0.01)$  , includes variability
- **Initial count  $N(0)$  and growth rate  $\mu$ ?**, given observation  $N(50)=18$ .  
Compute:  $p(N(0), \mu \mid N(50)=18)$ , includes uncertainty and variability



# Posterior distribution for $N(0)$ & $\mu$

- The joint distribution of  $N(0)$  and  $\mu$  shows there are some **probable combinations of values!**
- Most likely, either  $\mu$  was large and  $N(0)$  small,  
... or  $N(0)$  large and  $\mu$  small,  
... or some of the blue dots in the figure →
- This is also known as the *identifiability problem* of individual parameters.

Read more: <https://doi.org/10.1111/risa.13386>





# More observations: $N(t_1), N(t_2), \dots$

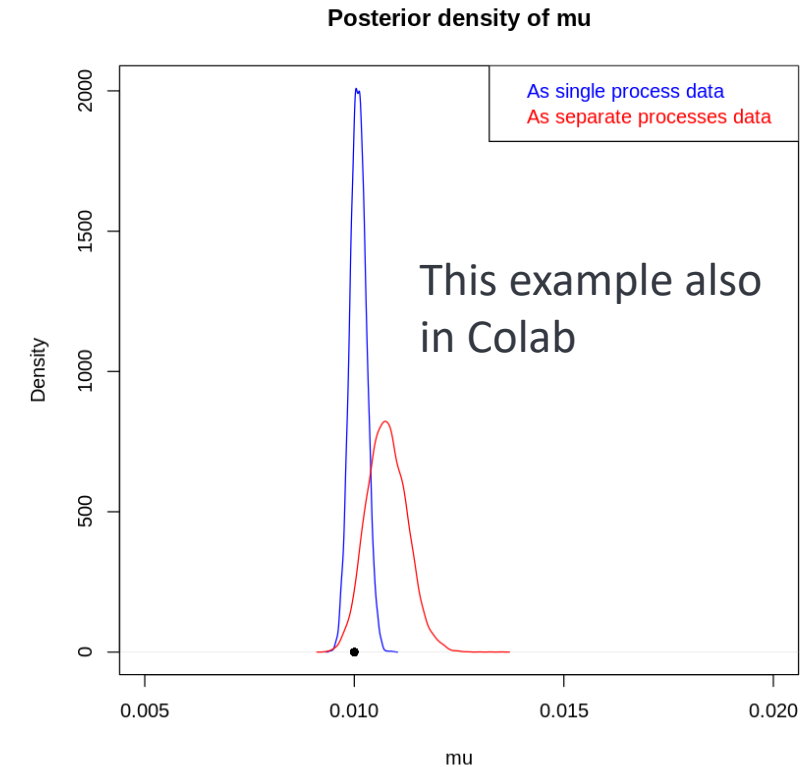
- Either data from the **same** growth process, or **separate** processes?

- If **separate** (independent) processes, the stochastic model for the data is:

$$P(\text{data}) = \prod_{i=2}^k p(N(t_i) \mid N(t_1), \mu) \quad N(t_i) \text{ depends on } \textit{initial state}.$$

- If the **same** process, then the stochastic model for the data is:

$$P(\text{data}) = \prod_{i=2}^k p(N(t_i) \mid N(t_{i-1}), \mu) \quad N(t_i) \text{ depends on } \textit{previous state}.$$





## *In Colab: estimation with counts $N(t_1), \dots, N(t_k)$*

- Assume observed counts from the same process, at times  $t_i$ :

- Probability for observed counts:**

$$\begin{aligned} p(\text{data} \mid \mu) &= \prod_{i=2}^k p(N(t_i) \mid N(t_{i-1}), \mu) \\ &= \prod_{i=2}^k \text{NegBin}(N(t_i) - N(t_{i-1}) \mid N(t_{i-1}), e^{-\mu(t_i - t_{i-1})}) \end{aligned}$$

- Prior probability for  $\mu$ :**  $p(\mu) = \text{Uniform}(\dots)$

- Posterior probability for  $\mu$ :  $p(\mu \mid \text{data}) = \prod_{i=2}^k p(N(t_i) \mid N(t_{i-1}), \mu) p(\mu) / \text{const.}$

Model  
for data

Prior

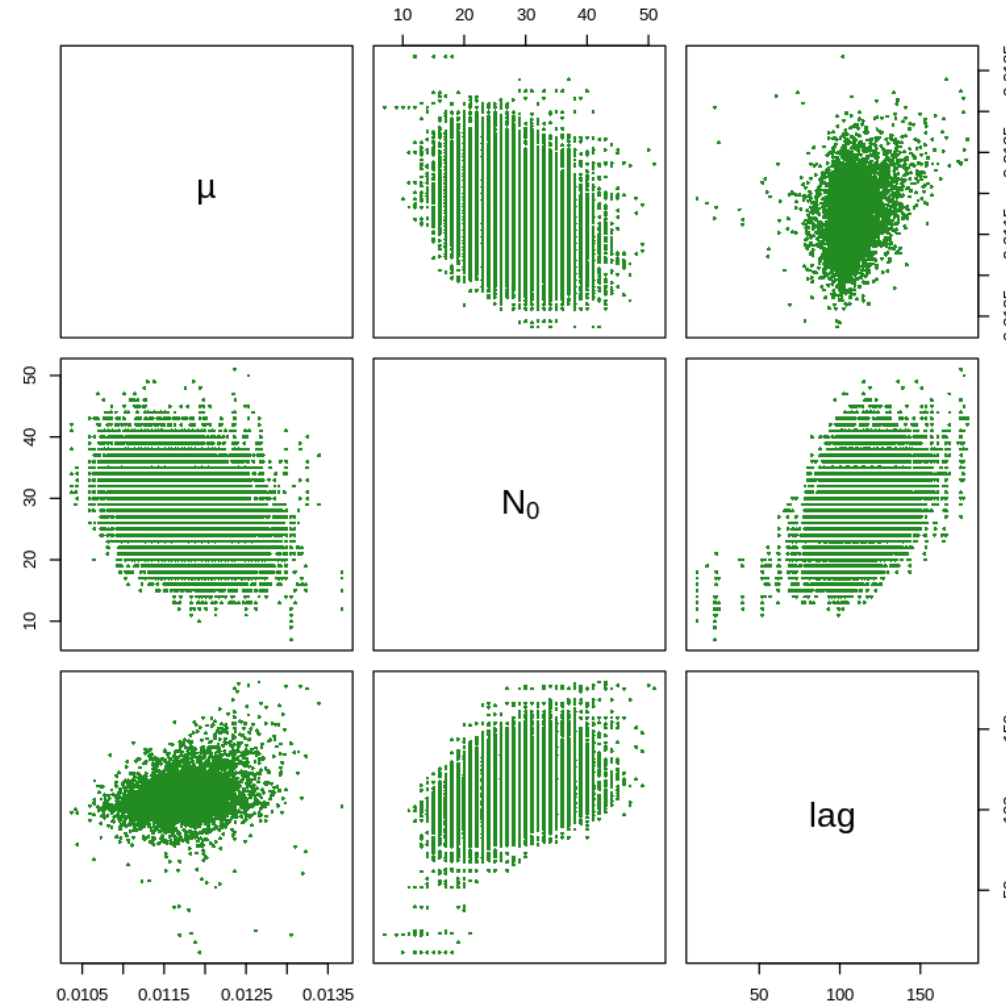




## *In Colab: with measurement (observation) error*

- Exact observations are not often possible, we need some **probability model** for the *observed measurements*.
  - In stochastic model:  $E( N_{\text{obs}}(t_i) ) = N(t_i)$  where  $N(t_i)$  is the true count, for which the stochastic model is written.
  - In deterministic model:  $E( N_{\text{obs}}(t_i) ) = f(t_i)$  where  $f(t_i)$  is the deterministic function. (it may correspond to the mean of the stochastic model).
  - Both situations need **observation model** with variance parameter to describe the deviations of observed values from the expected value.

# In Colab example: you might get something like this: (estimation of $\mu$ , $N_0$ , lag, all together)





# Modeling growth with log-counts

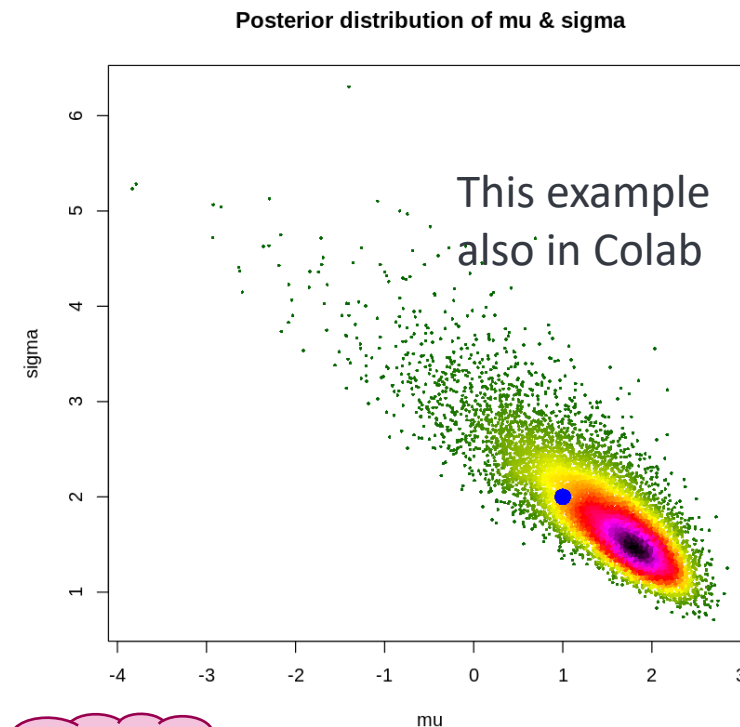
- Deterministic model for the mean:  $E(\log N(t)) = \log N(0) + \mu t$
- Observed measurements  $\log N(t) \sim \text{Normal}(\log N(0) + \mu t, \sigma^2)$ 
  - Mean function is logarithm of the stochastic process mean  $N(0)\exp(\mu t)$ .
  - Variance  $\sigma^2$  is a combination of both stochastic variation and observation error.
  - Bayesian inference for  $\log N(0)$ ,  $\mu$ ,  $\sigma$ , based on observed  $\log N(t_i)$

$$p(\mu, \sigma, \log N(0) \mid \text{data}) = \prod_{i=1}^k p(\log N(t_i) \mid E(\log N(t_i)), \sigma) p(\mu, \sigma, \log N(0)) / \text{const}$$



# Estimating distribution from partially left-censored data: principles

- More details in practical example code (Colab)!
  - For example: normal distribution for log-concentrations.
  - Limit of quantification LOQ
  - k data points >LOQ, m data points <LOQ



Exact data points      Left-censored data      Prior

$$p(\mu, \sigma \mid \text{data}) = \prod_{i=1}^n p(\log(c_i) \mid \mu, \sigma) \times F(\text{LOQ} \mid \mu, \sigma)^m \times p(\mu, \sigma) / \text{const}$$

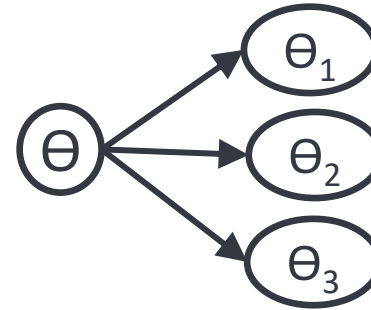
F is cumulative probability function at LOQ.



# Multilevel (hierarchical) models

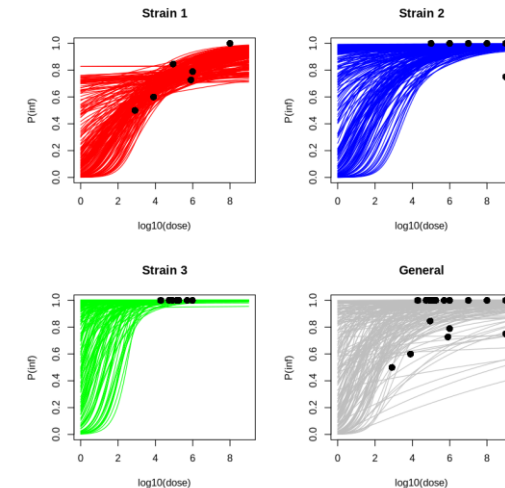
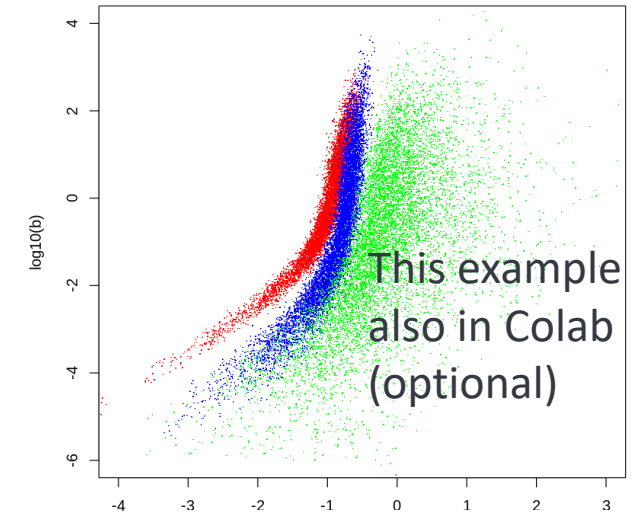
- More details in practical example codes (Colab)!

- Describe *several layers* of variations:
  - Between studies/outbreaks/experiments
  - Between strains
  - Within strains



- In dose-response models, from Teunis et al: DOI: [10.1017/S0950268807008771](https://doi.org/10.1017/S0950268807008771) & Anne Thebault (Example in this workshop!)
- In inactivation/growth models, e.g. Koyama et al: DOI: [10.1128/AEM.00918-21](https://doi.org/10.1128/AEM.00918-21)

Posteriors of DR-parameters for 3 strains





...Thank you!

Next part:

Summary of the tools & Colab.

Examples in Colab by Jukka, Anne, Hiroki, Kento



# Summary of the computational software

- **BUGS and JAGS** very similar, nearly same syntax for model definition.
  - Differences in parameterization of distributions.
  - Use different **R-packages** for calling BUGS/JAGS and processing outputs.
  - Use *Gibbs sampling algorithms* and other variants.
  - BUGS also has its own graphical UI, but only in Windows. It also allows building models "graphically" as DAGs (Directed Acyclic Graphs of the Bayesian model).
- **Stan** utilizes Hamiltonian MCMC method for simulating from the posterior distribution. Based on *derivatives* of functions of parameters.
  - Not good if discrete parameters to be sampled (derivatives not defined).
  - Easy to get posterior mode?
  - not necessary to specify priors (uses automatic priors)
  - Can use brms package, rstan.
- **All can be used in R (here in Colab!).**



# Common terminology:

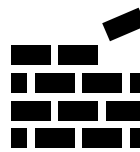
- For the tools, bayesian models are made of these "nodes":
  - **Constants** (given as assumptions, no model defined for these).
  - **Observed variables** (given as fixed values in data, a model is defined for these).
  - **Parameters and unobserved variables** (model or prior is defined for these).
  - Posterior distribution is computed for *all unobserved variables and parameters*.
    - Hence the Bayesian model is fully specified only after the data is specified!

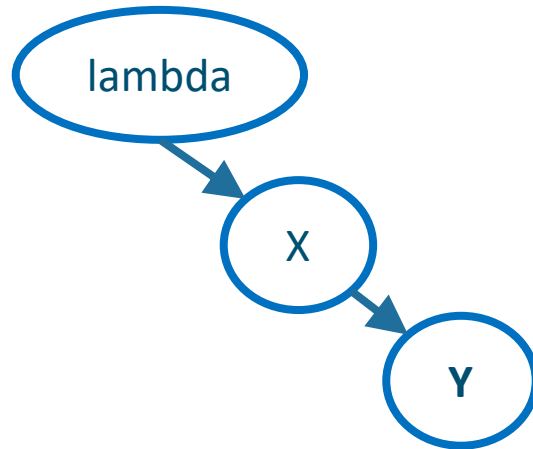




# Building Bayesian inference with the tools:

- You define the same *necessary blocks* in any of the tools BUGS/JAGS/Stan
  - Basically, very similar structure, but differences in syntax!

 You can think as **DAG** of the model:  
(Directed Acyclic Graph of conditional distributions)



in BUGS the same structure would be coded:

```
model{  
  for(i in 1:n){  
    Y[i] ~ dbin(p,X[i])    # P(data | parameter)  
    X[i] ~ dpois(lambda)  
  }  
  lambda ~ dgamma(0.01,0.01) # prior  
}
```

↑ This would define a Poisson model for variables  $X[i]$ , and a prior for the parameter  $\lambda$ . And a binomial model for data  $Y[i]$ .

The tools can then give you a random sample from the posterior  $p(\lambda, X | Y)$ .



# Technical issues: left-censored data, using BUGS/JAGS/Stan

- In BUGS/JAGS/Stan, the syntax differs, but in principle you need to define:
  - (1) Probability density for the exact measurements
  - (2A) Probability for the censored observations which will be imputed
    - BUGS:  $y[i] \sim \text{dnorm}(\mu, \tau) \text{C}(\text{upper}, \text{lower})$
    - JAGS:  $\text{observed}[i] \sim \text{dinterval}(y[i], \text{limits}[1:2])$
    - Stan: `array[N_cens] real<upper=LOQ> y_cens;`
  - (2B) Probability for the censoring event, without imputing censored values
    - Using Bernoulli model with cumulative probability  $F(\text{LOQ})$  in BUGS or JAGS  
`one[i] <- 1 ; one[i] ~ dbern(pr[i]) ; pr[i] <- phi( ( LOQ[i]-mu)*sqrt(tau) )`
    - Special syntax for likelihood expression in Stan.

Do you need only posterior distribution for parameters – or also imputation of the censored values? - both are possible to get!



# Downloads

- OpenBUGS
  - <https://www.mrc-bsu.cam.ac.uk/software/bugs/openbugs/>
  - Easy to start, versatile. Need r-package r2OpenBUGS.
  - Earlier: WinBUGS, new development: [MultiBUGS](#)
- JAGS
  - <https://sourceforge.net/projects/mcmc-jags/files/latest/download>
  - Need various packages rjags, runjags, coda, r2jags
- Stan
  - <https://mc-stan.org/users/interfaces/rstan>
  - More technical model specifications.
  - Brms interface to fit Bayesian generalized (non)linear multivariate multilevel models using Stan.



# With all tools

- Don't lose sight of what is the Bayesian model you try to make!
  - *Model* itself -vs- *implementation* of the model.
- Initial values can be critical, priors can be critical.
- MCMC may not converge well.
  - Numerical errors can occur, could crash.
  - The model is poorly identifiable – or not even properly defined.
  - Data may be simply too poor.
  - MCMC sampling algorithm may not be good.
  - Try multiple chains with different initial values.
- Compare the results with other estimation methods.
- Try with simple models first!

# Tutorial for google colaboratory





# What is Google Colaboratory?

- ◆ **Free Access:** Google Colab is free to use, providing an environment to write and **execute code for Python and R.**
- ◆ **Interactive Coding Environment:** Google Colab provides an interactive environment like Jupyter Notebooks where you can mix **text, code, and outputs all in one document.**
- ◆ **Sharing and Collaboration:** Google Colab is integrated with Google Drive, allowing you to share notebooks with other users and collaborate on work.



# Limitation of Google Colaboratory

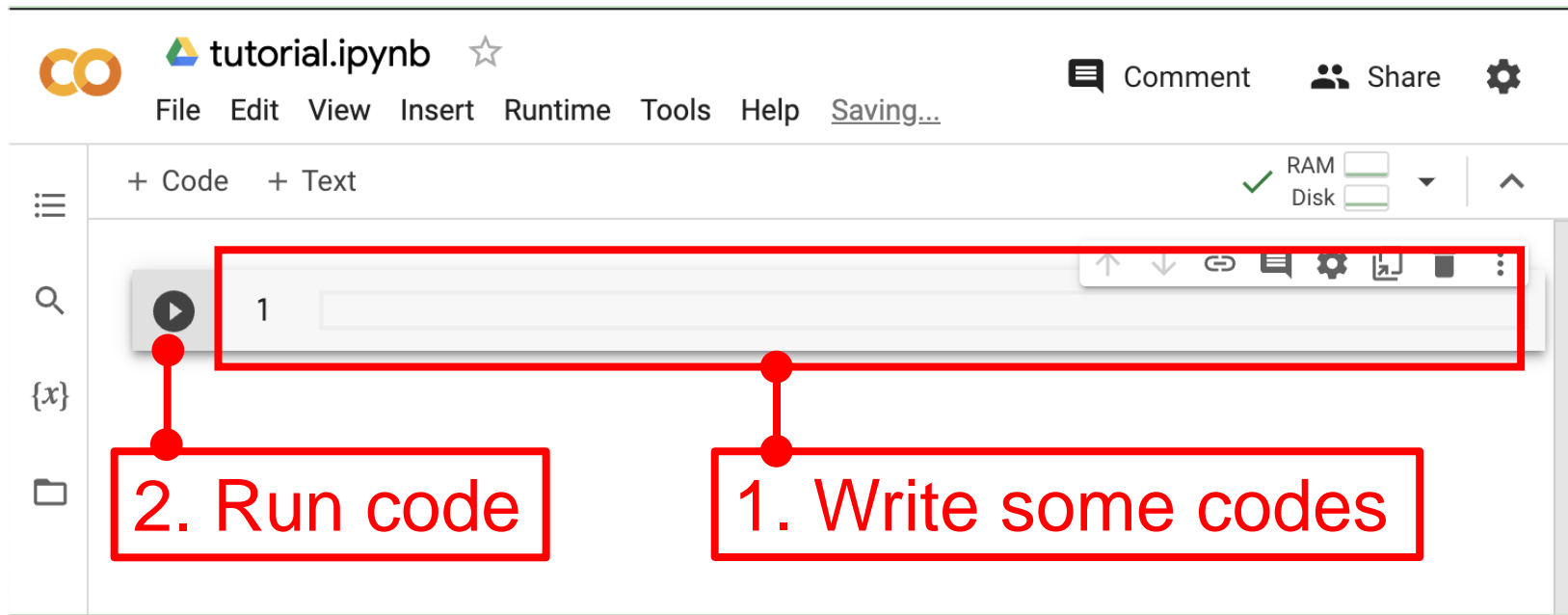
- ◆ **Limited Resource Allocation:** The usage of resources are limited. High-demand workloads may face interruptions.
- ◆ **Privacy Concerns:** Because notebooks are stored on Google Drive, data security and privacy may be a concern, especially for sensitive datasets.
- ◆ **File Storage:** While Google Colab integrates well with Google Drive, working with large datasets can be challenging due to the limitations of Drive's storage

# How to use Google Colaboratory for Python





# Running codes



tutorial.ipynb ☆

File Edit View Insert Runtime Tools Help Saving...

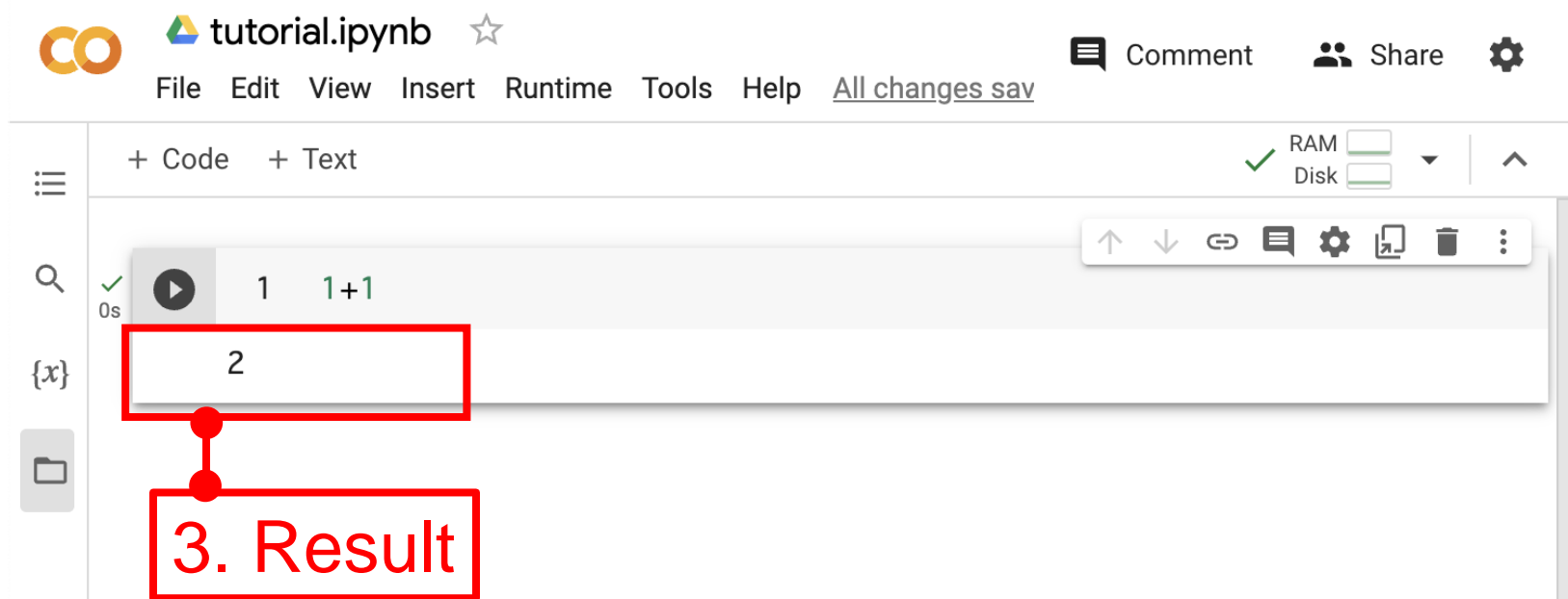
+ Code + Text

RAM ☐ Disk ☐

1

2. Run code

1. Write some codes



tutorial.ipynb ☆

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

RAM ☐ Disk ☐

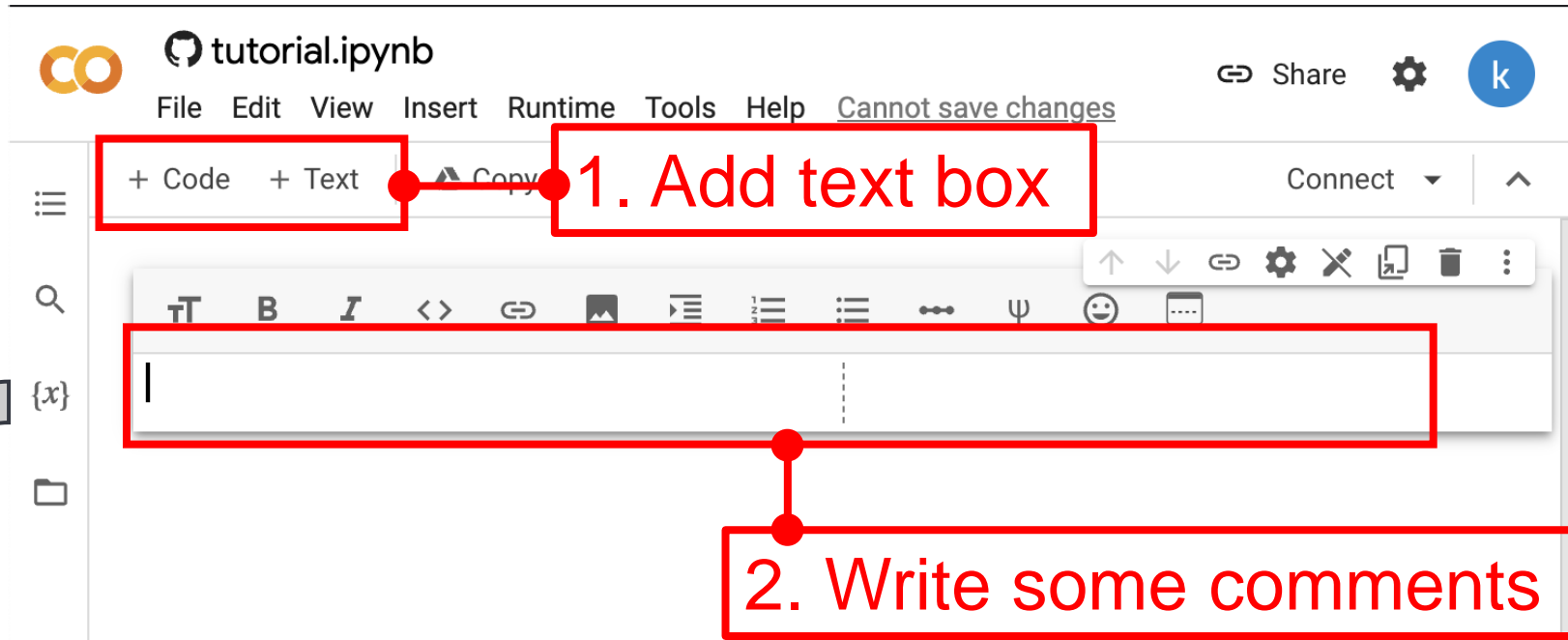
✓ 0s 1 1+1

2

3. Result



# Writing note



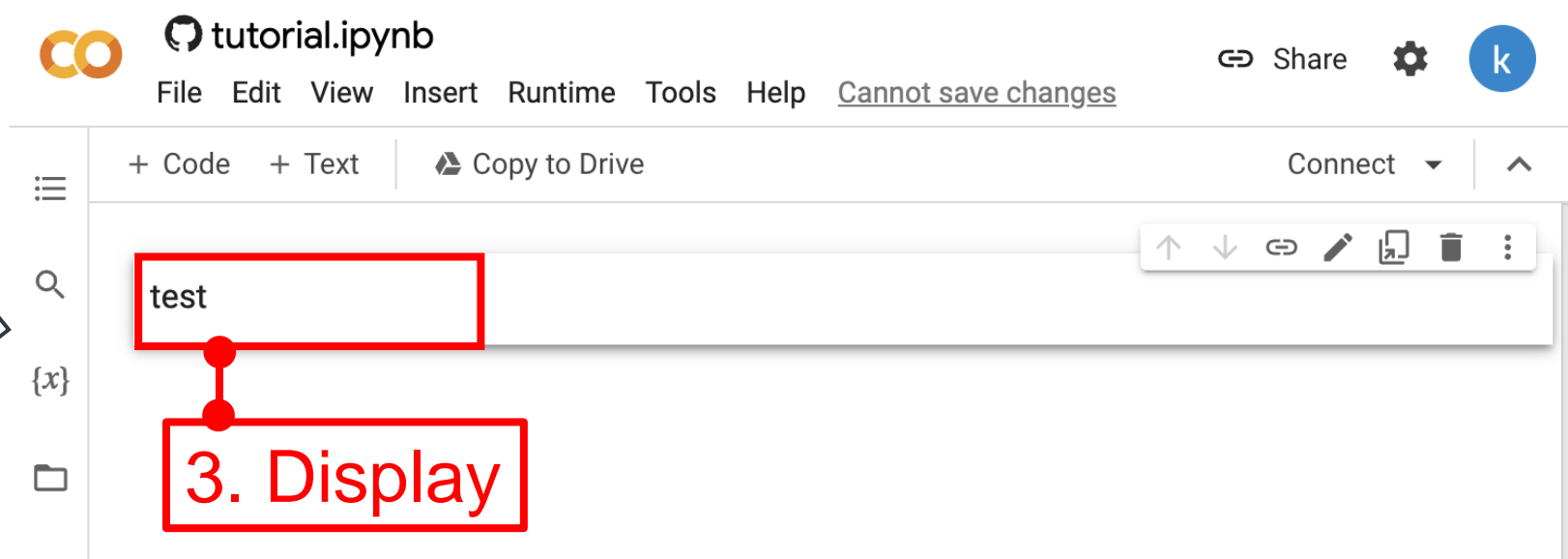
tutorial.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy

1. Add text box

2. Write some comments



tutorial.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

test

3. Display



# Uploading and downloading files

The screenshot displays the JupyterLab interface with two panels. The top panel shows a code editor with a cell containing the code `1 1+1` and its output `2`. The bottom panel shows a file browser with a folder named `sample_data`. Red annotations highlight key file management features:

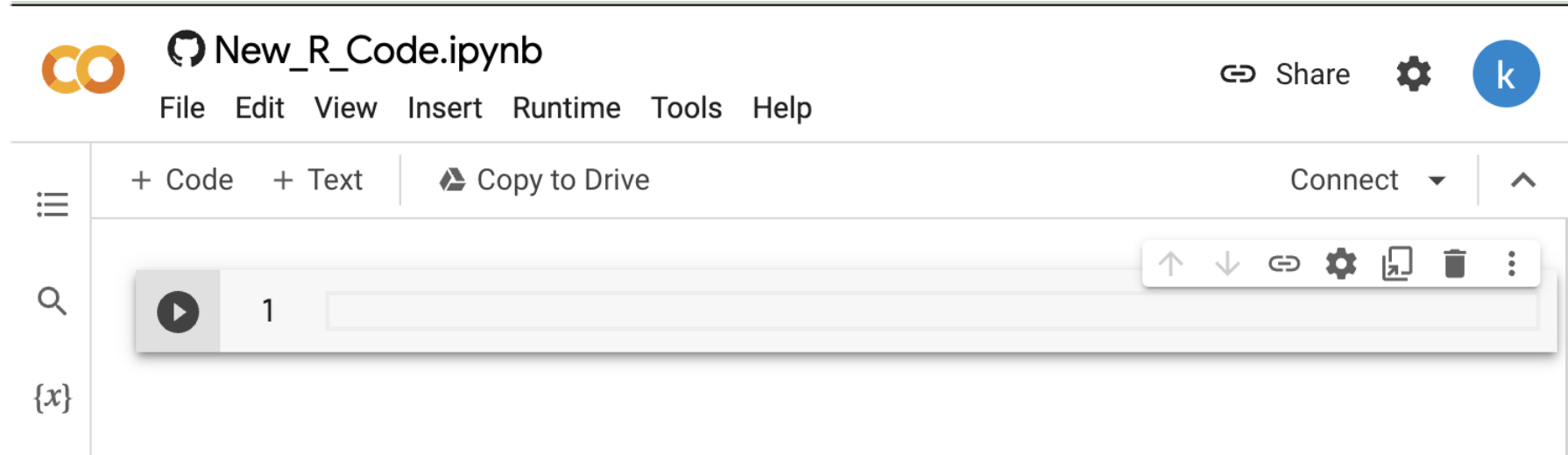
- Open folder:** A red box highlights the folder icon in the left sidebar, with an arrow pointing to it from the label.
- Upload files by drop-and-dragging:** A red box highlights the `sample_data` folder in the file browser, with an arrow pointing to it from the label.
- Download files:** A red box highlights the download icon (a downward arrow) in the file browser, with an arrow pointing to it from the label.

A large curved arrow on the left side of the image points from the folder icon in the top panel to the file browser in the bottom panel, indicating the transition between the two views.

# How to use Google Colaboratory for R



# Running code



- ◇ The use of notebook is the same as python.
- ◇ R environment setting is needed. We distribute the environment R through the following links:  
[https://github.com/kento-koyama/bayesian\\_predictive\\_micro\\_ICPMF12/blob/main/Extra\\_templates/New\\_R\\_Code.ipynb](https://github.com/kento-koyama/bayesian_predictive_micro_ICPMF12/blob/main/Extra_templates/New_R_Code.ipynb)



**anses**

- **EXAMPLE 1: CENSORED DATA (PARTICULAR CASE OF MISSING DATA)**

# 1 — Censored data



## General considerations

**General : censored data are missing data with an information (LOD, LOQ, censored data)**

**Missing data can appear in the outcome variable (prediction on a new data set ) or in the predictors ;**

**Data can be heterogeneous: predictor can be partially known in some parts of the dataset;**

**Application : Observational studies, meta-analysis, epidemiological studies, transfer parameters...**

**In frequentist approach:**

No covariate: « fitdistrplus » package

multivariate analysis : limitation to complete dataset (suppressing lines of missing values)  
; or imputation: package « mice » in R...



## General considerations

Estimate parameters of a censored distribution? `fitdistrplus` package in R is doing so well

Estimate effect of predictors? `Survreg` package in R is also working (but no imputation)

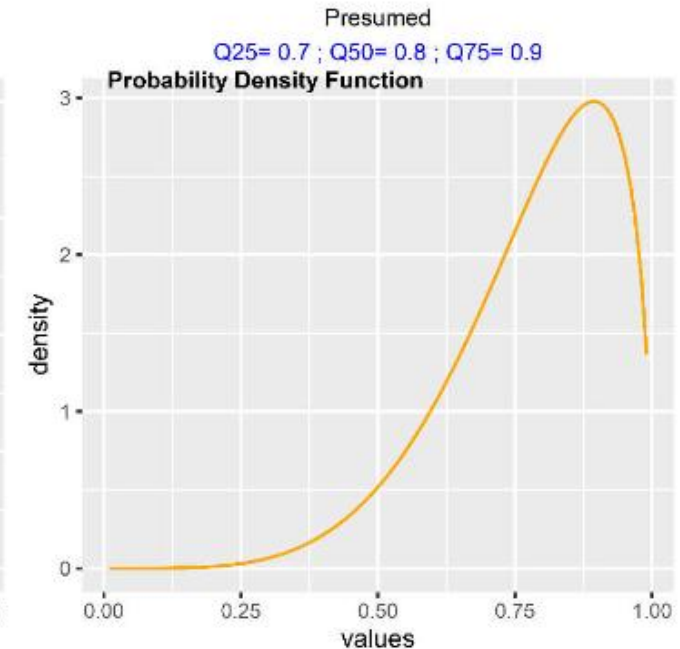
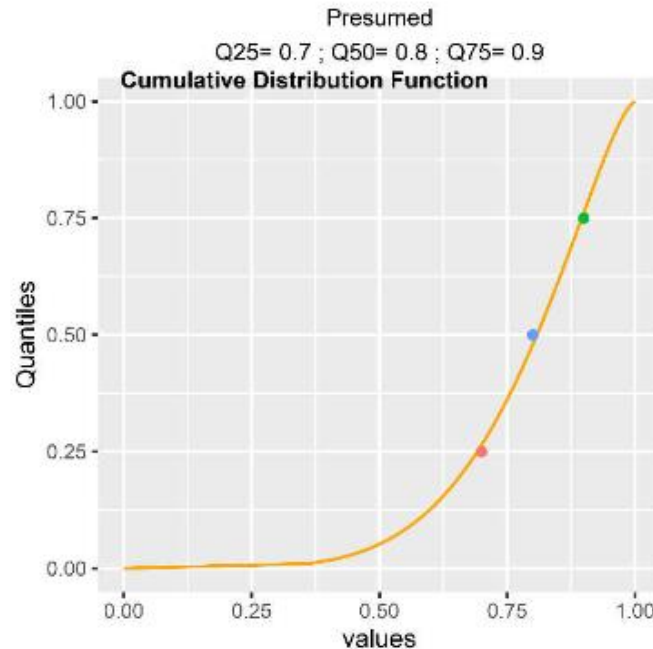
If you need imputation of censored values: to use those imputed data (low contamination values close to LOQ(norovirus), outbreaks data, sum of censored in toxins/chemicals...

And even so new packages are taking into account censoring imputation (not tested)

But always limitations as all predictors follow *a Normal distribution....or linear relationship or a lot of data using a « black box » method...* Bayesian approach is giving more flexibility without a priori limitations...

# Principle

- **Replacement/imputation of missing values**
- **No replacement nor input**
  - Use of CDF for censored value
  - pdf for non censored values
- **Adding covariates is giving information**



# 1 — Dose-response





# (mechanistic) Dose-Response models

- First back to concept:
- For infection two sequential process are assumed to occur (Haas *et al.*, 1999):
- “The human host must ingest one or more organisms that are capable of causing infection or disease”. This is the exposure step, and the probability to ingest  $j$  organisms, knowing the mean dose of exposure,  $d$ . The notation is  $P1(j/d)$
- “(...) only a fraction of the ingested organisms reach a site where infection can begin”. The probability of  $k$  organisms to infect, knowing  $j$  are ingested is noted  $P2(k/j)$

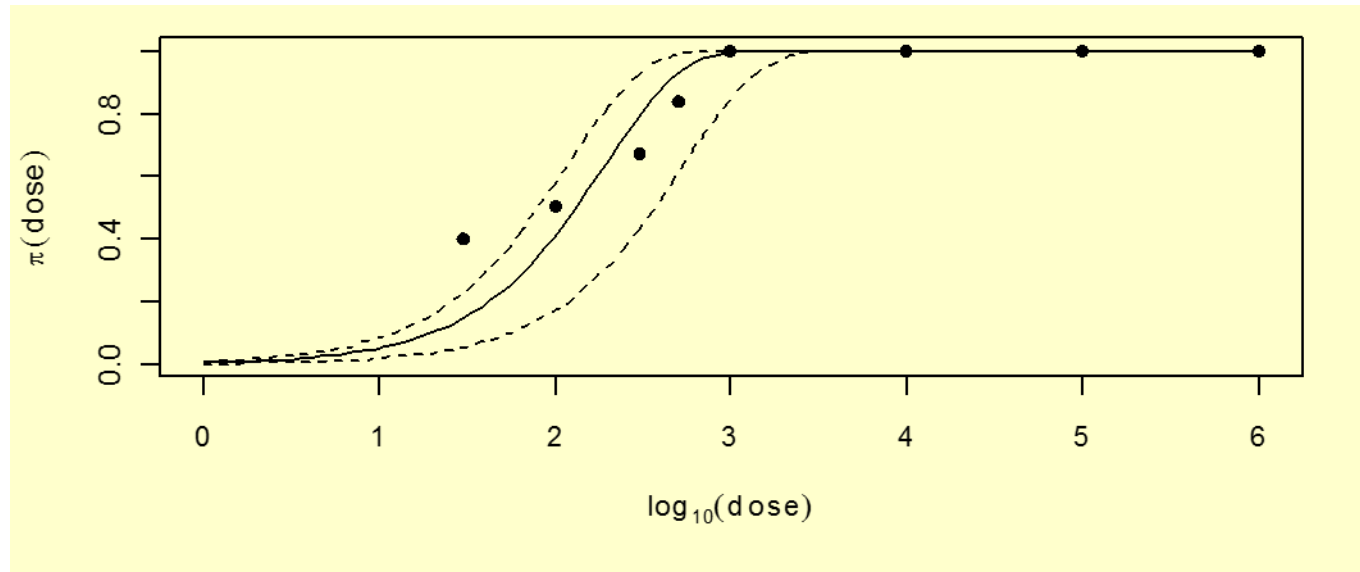
$$P \inf(\bar{d}) = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} P1(j / \bar{d}) P2(k / j)$$



## exponential dose-response

- Knowing that  $p_m$  is a probability that one infectious particle infect,  $1-p_m$  the probability that is this virus not infect, and for  $n$  virus ingested,  $P_{inf}$  is the probability that at least one virus succeeds in infecting the exposed individual. If  $p_m$  is constant
- For an individual  $i$ :
- $n[i, \lambda] \sim \text{Poisson}([\lambda])$
- $P_{inf}[i, \lambda] = 1 - (1 - p_m)^{n[i, \lambda]}$
- $inf[i] \sim \text{Bernoulli}(P_{inf}[i, \lambda])$
- The marginal risk of infection is a function of  $\lambda$  if  $p$  constant and dose Poisson 
$$P_{inf}_m = 1 - \exp(-\lambda \times p_m)$$

# Ex Cryptosporidium, Toxoplasma...



Low dose  
extrapolation



# Beta Poisson model

- $p_m \sim \text{Beta}(\alpha, \beta)$
- $\text{dose} \sim \text{Poisson}(\lambda)$
- whenever  $\beta \gg 1$  and  $\alpha < \beta$  can be approximated by the equation (Furumoto, et Mickey, 1967)

$$P_{\text{inf}/\lambda} = 1 - \left(1 + \frac{\lambda}{\beta}\right)^{-\alpha} \text{ (not single hit low dose)}$$

$$DMI_{50} = \beta(2^{1/\alpha} - 1)$$

- If not  $\beta \gg 1$  and  $\alpha < \beta$ :  ${}_1F_1$  (Kummer confluent hypergeometric function)(Poisson distribution)

$$P_I(\lambda_k) = 1 - {}_1F_1(\alpha, \alpha + \beta, -\lambda_k)$$

# Dose-response (classical Beta-Binomial)



The general equation of Betabinomial model, assuming  $p_{inf}$  (survival of each pathogen) can be modelled with a Beta distribution

$p_m \sim \text{Beta}(\alpha, \beta)$

With  $n \sim \text{NegBin}(\mu, \text{size})$

$P_m \sim \text{Beta}(\alpha, \beta)$

Individual  $i$ , particle  $j$ ,

And  $\lambda$  is a function of  $\mu$  and  $\text{size}$  (Poisson

Gamma or NegBin)

$$P_{inf}[i, \lambda] = 1 - \prod_{j=1}^{n[i, \mu, \text{size}]} (1 - p_m[j])$$

$$P_{inf}(d / \alpha, \beta) = 1 - \frac{\Gamma(\alpha + \beta) \Gamma(\beta + d)}{\Gamma(\beta) \Gamma(\alpha + \beta + d)}$$

Where  $\Gamma$  is a Euler gamma function

D dose (can be modelled as Poisson, Negative Binomiale)





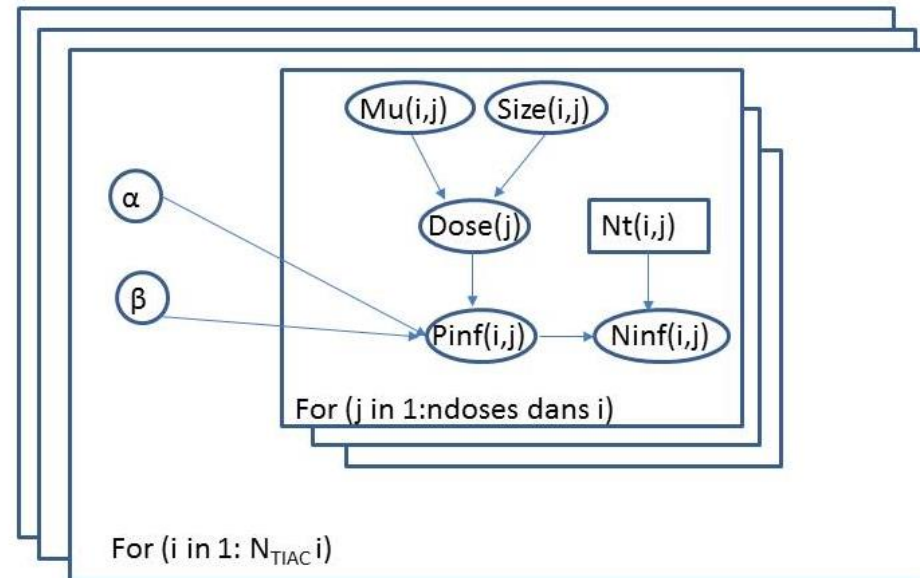
# Beta-Binomial general formulation

- If we simulate  $N$  (with enough iterations), mean of  $P_{\text{inf}}$  (integrate or  $\Sigma$ )
- Poisson is simulated  $P_{\text{inf}}$  is equivalent (1F1 with Poisson)
- Approximation of the integral is feasible if we simulate  $d$  (Poisson or NegBin or..)

$$P_{\text{inf}}(d / \alpha, \beta) = 1 - \frac{\Gamma(\alpha + \beta) \Gamma(\beta + d)}{\Gamma(\beta) \Gamma(\alpha + \beta + d)}$$

Use of gamma function in bayesian dose-response framework first described in Thebault et al., 2013

# DAG of betaBinomial hierarchical model with NegBinomial dose





# Trichinella hierarchical dose-response model

- We need one male and one female at least to obtain infectious larvae
- If the males and the females have an equal probability of survival **pm**,
- proportion of females is **r** and of males **1-r**,
- The probability of infection after having ingested N parasites ( $N_{\text{males}} + N_{\text{females}}$ ) (Takumi et al., 2009; Teunis et al., 2012) is :

$$P_{\text{inf}} = \left[ 1 - (1 - p_m)^{N_{\text{males}}} \right] \times \left[ 1 - (1 - p_m)^{N_{\text{females}}} \right] = 1 + (1 - p_m)^N - (1 - p_m)^{r \times N} - (1 - p_m)^{(1-r) \times N}$$

And now we have  
with the Gamma  
function :

$$P_{\text{inf}}(d / \alpha, \beta) = 1 + \frac{\Gamma(\alpha + \beta) \Gamma(\beta + N)}{\Gamma(\beta) \Gamma(\alpha + \beta + N)} - \frac{\Gamma(\alpha + \beta) \Gamma(\beta + N_F)}{\Gamma(\beta) \Gamma(\alpha + \beta + N_F)} - \frac{\Gamma(\alpha + \beta) \Gamma(\beta + N_M)}{\Gamma(\beta) \Gamma(\alpha + \beta + N_M)}$$



# Paramerization of hyperparameters of Beta distribution $\alpha$ and $\beta$ (Teunis,2012)

$$E(X) = \frac{\alpha}{\alpha + \beta} = u$$



$$\alpha = u \left( \frac{u(1-u)}{\text{var}} - 1 \right) = u \times v$$

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} = v$$

$$\beta = (1-u) \left( \frac{u(1-u)}{\text{var}} - 1 \right) = (1-u) \times v$$

$w = \text{logit}(u)$  and  $z = \log(v)$

Can be also written in the form:

$u = \exp(w) / (1 + \exp(w))$  et  $v = \exp(z)$ .

$w \sim \text{Normale}(\text{meanw}, \text{tw})$

$z \sim \text{Normale}(\text{meanz}, \text{tz})$



For Normal distribution

$(\tau = 1/(\text{sd}^2))$

$\text{meanw} \sim \text{Normale}(-2, 0.5)$

$\text{meanz} \sim \text{Normale}(2, 0.5)$

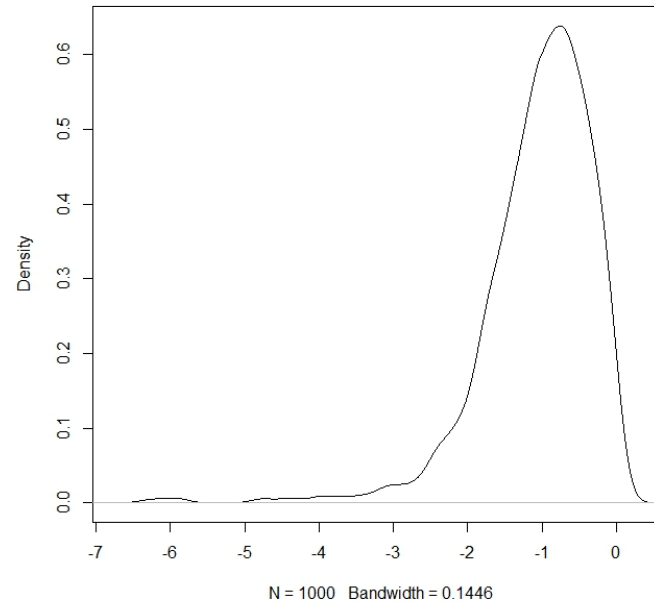
$\text{tw} \sim \text{Gamma}(0.5, 0.5)$

$\text{tdz} \sim \text{Gamma}(0.5, 0.5)$



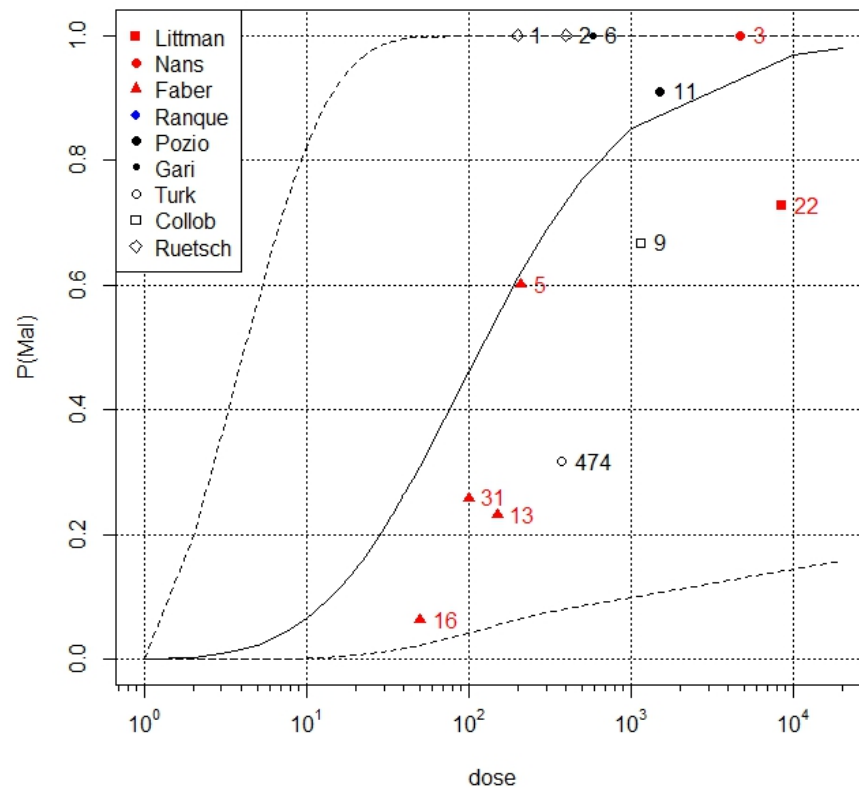
# Prior of the mean of Beta

On 1000 iterations  
min  $10^{-7}$ , max  
0.9999994, mean 0.2



Log10 scale

# Results hierarchical model with outbreaks (Anses,2016)



Hierarchical :

1 value of Beta parameters for **each outbreak**

Those values are taken **from a distribution with hyper-parameters**

The overall shape is for all outbreaks  
Same idea with random effect in mixed models...but applied to alpha and beta parameters

The hyperparameters are fitted for outbreak level (number of outbreaks is important); same approach can be applied to strains...



# Post –hoc controls in rjags

## Convergence :

- The Gelman and Rubin criterion compares the intra- and inter-chain variances of the posterior distributions and must tend to 1
- The posterior distributions must be Normal (<http://127.0.0.1:15831/library/coda/html/gelman.diag.html>).
- The median value of the PRSF (Potential Scale Reduction Factor) must be less than 1.1 and close to 1 to be acceptable.
- The upper value of the 95% credibility interval of PRSF must in this case also be less than 1.1.
- The distributions of the parameters  $\alpha$  and  $\beta$  do not follow a Normal law.
- Their reparameterization in  $z$  and  $w$  follows a quasi-normal law.
- It is therefore on these two parameters that the criterion of Gelman and Rubin will be applied.



# Post –hoc controls in rjags

## Correlations between parameters

This criterion must highlight the non-existence of correlations (level lower than 0.05) when they are not expected.

Correlations between parameters  $\alpha$  and  $\beta$  are expected (average risk of infection).

**Autocorrelations** Despite the achievement of the stationarity phase, autocorrelations may remain in the a posteriori sample.

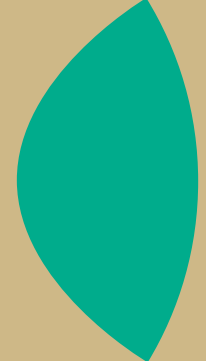
The thin parameter corresponds to the sampling step in the posterior distribution allowing to keep in the posterior sample only uncorrelated values at the threshold of 0.05.

## Quality fit

The estimation of the estimated DIC/BIC in jags is not applicable to hierarchical models. In a simplified way, the 95% credibility interval of the prediction must contain the observations (or the average of these) and will be visualized graphically. is (Takumi et al., 2009; Teunis et al., 2012)



# BIBLIOGRAPHY



## BIBLIOGRAPHY

### CENSORED DATA

**Comparison of methods EFSA :**

**<https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2010.1557>**

**Example : brms use for oyster norovirus contamination analysis :**

**<https://www.efsa.europa.eu/fr/efsajournal/pub/5762>**

**reference book for rstan : [https://mc-stan.org/docs/2\\_18/stan-users-guide/index.html](https://mc-stan.org/docs/2_18/stan-users-guide/index.html)**

Suzuki Y, Tanaka N, Akiyama H. Attempt of Bayesian Estimation from Left-censored Data Using the Markov Chain Monte Carlo Method: Exploring Cr(VI) Concentrations in Mineral Water Products. *Food Saf (Tokyo)*. 2020;8(4):67-89. Published 2020 Dec 25. doi:10.14252/foodsafetyfscj.D-20-00007

**Reference publication for rjags :** On Bayesian modeling of censored data in JAGS: 2022 Xinyue Qi, Shouhao Zhou and Martyn Plummer<sup>58</sup> BMC bioinformatics

# DOSE-RESPONSE



- Haas CN, Rose JB, Gerba CP. Quantitative microbial risk assessment. John Wiley and sons ed. New York: John Wiley and sons, USA; 1999.
- Isabelle Villena (1) , (..) Anne Thébault (18) Avis du 16 décembre 2016 révisé le 14 mars 2017 de l'Anses relatif à la contamination de produits de charcuterie crue par *Trichinella* spp.  
<https://hal.inrae.fr/hal-02796226>
- Takumi K, Teunis P, Fonville M, Vallee I, Boireau P, Nockler K, et al. Transmission risk of human trichinellosis. *Vet Parasitol.* 2009 Feb 23;159(3-4):324-7.
- Teunis PF, Nagelkerke NJD, Haas CN. Dose Response models for infectious gastroenteritis. *Risk Anal.* 1999;19(6):1251-9.
- Teunis PF, Koningstein M, Takumi K, van der Giessen JW. Human beings are highly susceptible to low doses of *Trichinella* spp. *Epidemiol Infect.* 2012 Feb;140(2):210-8.
- Thébault A. Probabilistic models for Risk assessment of viral infection associated with contaminated food consumption : public health impact of norovirus or hepatitis A virus contamination in oysters (dose-response chapter) 2013 in English-PhD thesis  
<https://pastel.archives-ouvertes.fr/tel-03560036>
- **Thébault A**, P.F.M. Teunis, J Le Pendu, F S. Le Guyader, J-B Denis. Infectivity of GI and GII noroviruses established from oyster related outbreaks (2013). *Epidemics* 5: 98-110.  
<http://www.sciencedirect.com/science/article/pii/S1755436512000576>.



**RUOKAVIRASTO**  
Livsmedelsverket • Finnish Food Authority

---

# Your turn for the conclusion

<https://app.klaxoon.com/join/7DUYWBN>