# $n = 3$

## Parameter inference with few experiments for a basic reaction network

Benjamin GAUDIN - Julien JOACHIM

Supervised by Gabriel LE BOULDER, Matthias FUEGGER and Thomas NOWAK

January 15, 2024

# The $n = 3$ problem

## Context

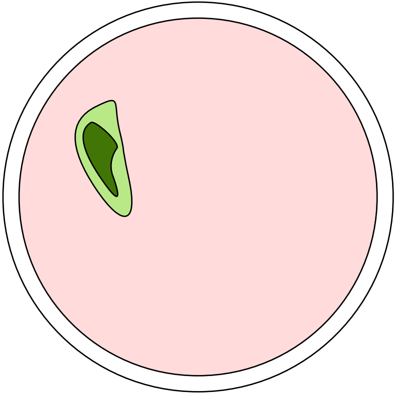


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem

## Context

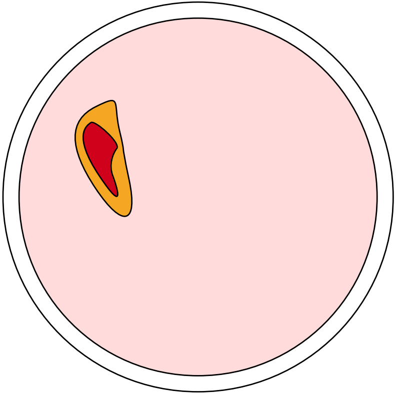


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem
## Context

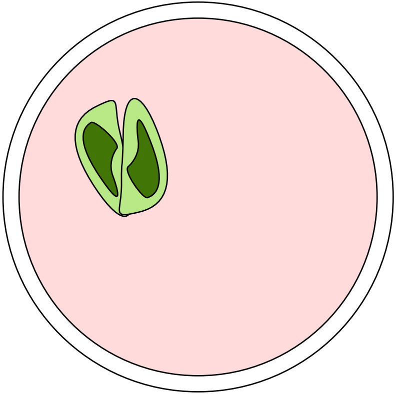


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem

## Context

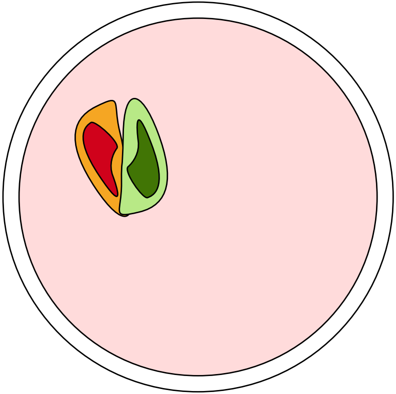


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem

## Context

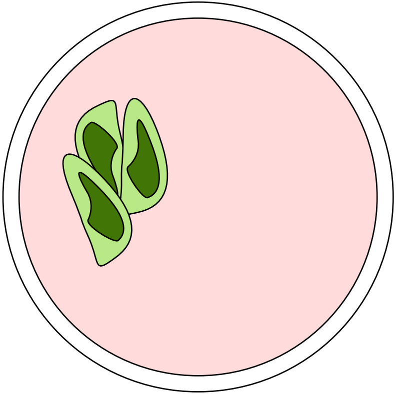


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem

## Context

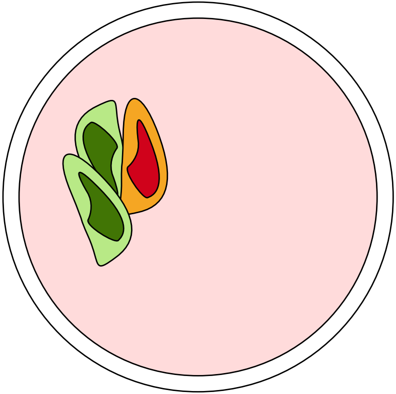


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem
## Context

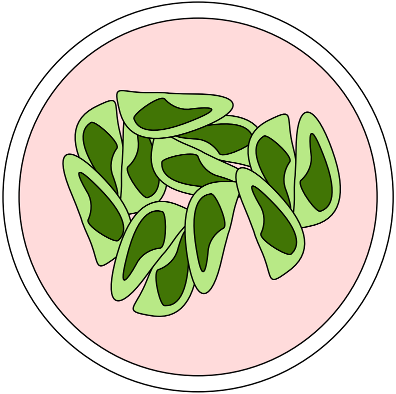


Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The $n = 3$ problem

## Context

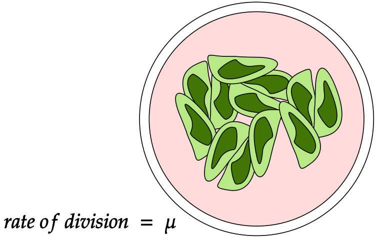



Figure: Petri dish with $k = 1$ bacterium

Experiment : Observe successive bacteria divisions over time

# The n=3 Problem

## Question

How to estimate $\mu$ reliably with as few experiments as possible? How many experiments are required to have a 95% confidence in our estimation?

# The $n = 3$ problem

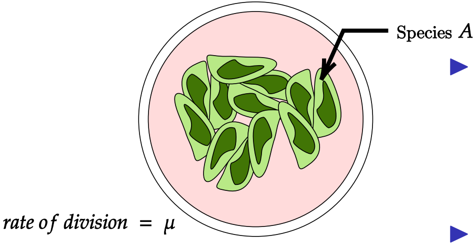



Figure: Petri dish with $k = 1$ bacterium

- System can be modeled by a *Reaction Network*

$$A \xrightarrow{\mu} 2A$$

- $\mu$ is the rate parameter : depends on the environment

# The $n = 3$ problem

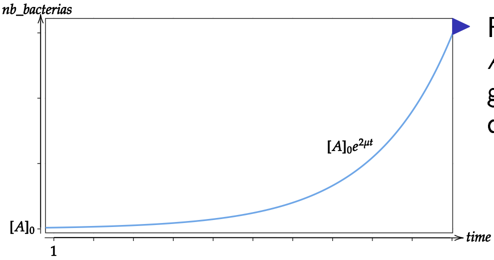



Figure: Petri dish with $k = 1$ bacterium

From the reaction network $A \xrightarrow{\mu} 2A$, we can derive a general expression for the quantity of cells:

- $\frac{d[A]}{dt} = 2\mu [A] \implies [A](t) = [A](0)e^{2\mu t}$
- so $\mu = \frac{1}{2t} \ln \left( \frac{1}{A_0} A(T) \right)$
- Doesn't work because it now a deterministic process

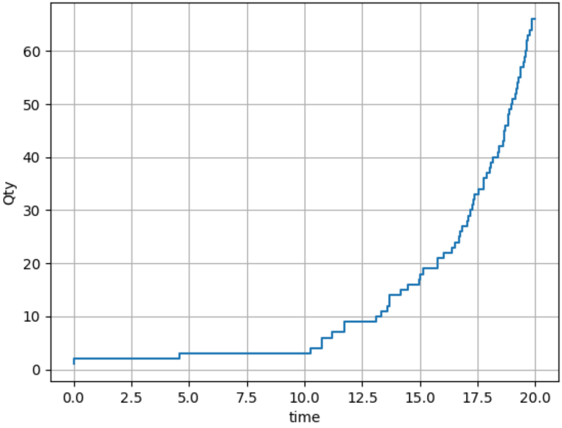
# Ideas explored

- Mathematical brute-force
- Using Cross-Entropy method

# New Idea (Thanks Matthias!:D)

## Simplifying assumption

We now assume that we have access to the whole trace of an experiment: we know the number of cells in the dish at every moment $t$



# New Idea

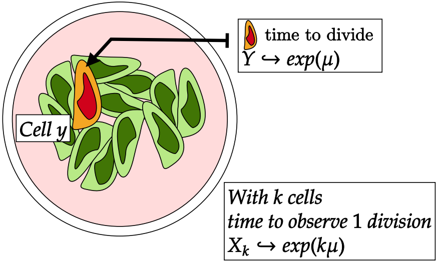
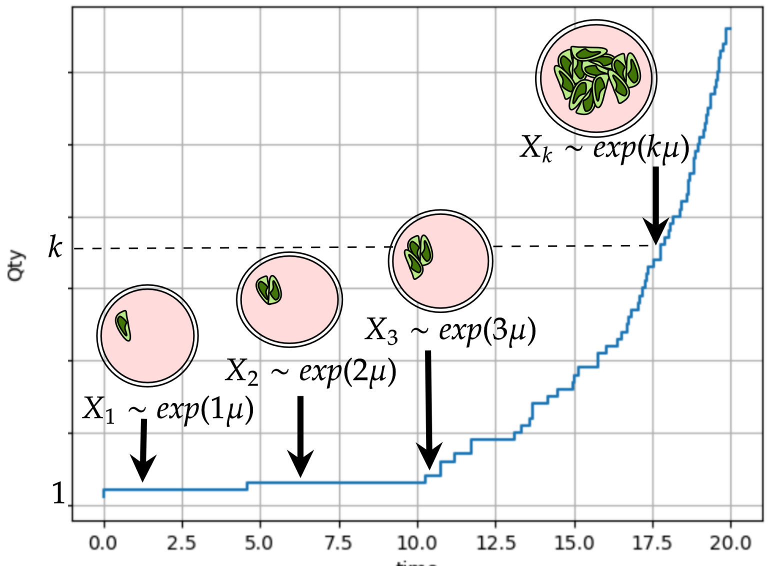



Figure: Petri dish with $k = 1$ bacterium

- When number of cell is large, the time that it takes for a each single cell to divide follows $\exp(\mu t)$
- So if there are $k$ cells in the dish, the wait time follows $\exp(k\mu t)$

# Modeling waiting time



# Modeling waiting time

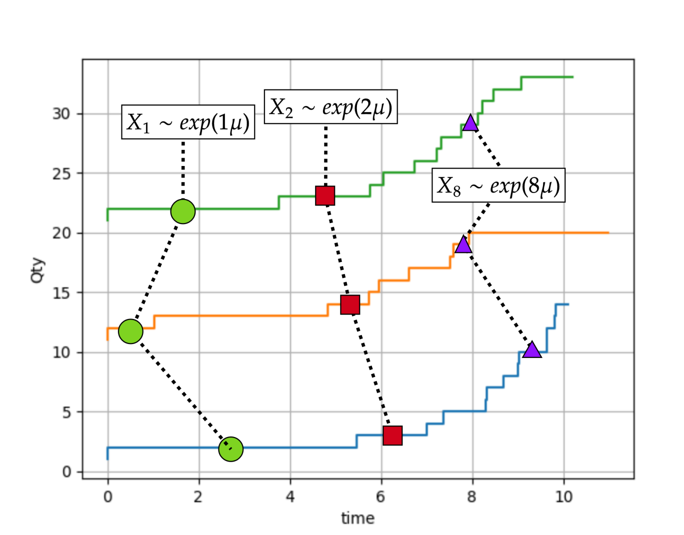



Figure: Petri dish with $k = 1$ bacterium

# Estimating $\mu$ with Maximum Likelihood

- If given a set of samples $X = (x_1, ..., x_n)$ drawn from a random variable $X_\lambda \hookrightarrow \exp(\lambda)$, we can estimate $\lambda$ with maximum likelihood by

$$\hat{\lambda} := MLE(X) = \arg\max_{\lambda > 0} \prod_{x_i} \frac{1}{\lambda} e^{\lambda x_i}$$

- Confidence : $\sqrt{n}\left|\lambda - \hat{\lambda}\right| \longrightarrow \mathcal{N}(0, some\_cov\_matrix^*)^1$
- So the bias $B = \mathbb{E}\left(\lambda - \hat{\lambda}\right) = o\left(\frac{1}{\sqrt{n}}\right)$
- $n = 3$? ☹

[1] The inverse of the Fischer Information Matrix

# A solution: Bootstrapping

- Technique for data re-sampling, when true distribution is unknown
- Assumption: Consider the observed data as representative of the true distribution $F$
- If $F$ is the real distribution, $\hat{F}$ is the empirical distribution :
  - $X \hookrightarrow \hat{F}$ means $P(X = x_i) = \dfrac{1}{n}$
- Idea : Do stats with $\hat{F}$ as we would with $F$

# Illustration of the Bootstrapping process

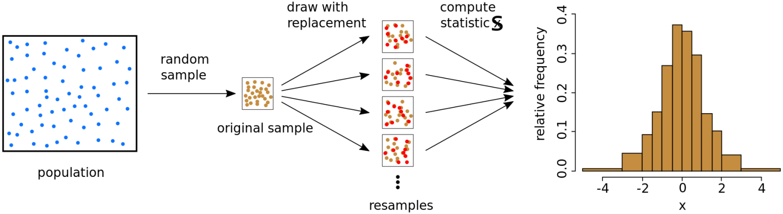



Figure: Petri dish with $k = 1$ bacterium

# WARNING

**WARNING : Assumption to use Bootstrapping**

We consider the observed data as representative of the true distribution $F$

- How to quantify the "closeness" of the observed distribution to the real one?
- Introduce a metric: *Statistical distance*
- $$\Delta(X, Y) = \int_{\mathcal{D}} P(X = \omega) - P(Y = \omega)\, d\omega$$
  - Need to be thought through...
- Assume for now that the assumption is satisfied...

# Bootstrap confidence interval of $\hat{\mu} = MLE(X)$

- We want a confidence interval on the value of $\mu$, $\hat{\mu}$
- Need to estimate the *Standard Error* on our Bootstrapped samples

# Bootstrap confidence interval of $\hat{\mu} = MLE(X)$

## Estimatino of the Standard Errroor

- Double Bootstrap : let $X_1^*, ..., X_R^*$ be $R$ replications of our original sample $y_1, ..., y_n \hookrightarrow exp(\mu)$
- Bootstrap Replicate $B$ times each $X_i^*$: Gives $R \times B$ samples $X^{**} = X_{1,1}^{**}, ..., X_{1,Q}^{**}, X_{2,1}^{**}, ..., X_{2,Q}^{**}, ..., X_{R,1}^{**}, ..., X_{R,Q}^{**}$.

$$\widehat{se}^2 = \frac{1}{n-1} \sum \left( \hat{\mu}\left(X_i^{**}\right) - \overline{\widehat{\mu}\left(X_i^{**}\right)} \right)^2$$

- $\text{var}(\widehat{se}) = \frac{O(1)}{n^2} + \frac{O(1)}{nB}$

# Bootstrap confidence interval of $\hat{\mu} = MLE(X)$

## Conf. Interv. Bounds

- We want a $1 - \alpha$ confidence interval on $\widehat{\mu}$,that is, $a$ and $b$ st
- $$P\left(a < |\widehat{\mu} - \mu| < b\right) = 1 - \alpha$$
- Idea : Studentize $\widehat{\mu}$

# Bootstrap confidence interval of $\hat{\mu} = MLE(X)$

## Conf. Interv. Bounds → Studentized variable

- $$Z = \frac{\widehat{\mu} - \mu}{\sigma_{\widehat{\mu}}} \equiv \frac{\widehat{\mu}^* - \widehat{\mu}}{\widehat{se}} \hookrightarrow \mathcal{N}(0,1)$$

- So after computations[2], we get a $1-2\alpha$ confidence interval for $\widehat{\mu}$:

$$low = \widehat{\mu} - \widehat{se}\cdot z^*_{(1-\alpha)(R+1)} \; high = \widehat{\mu} - \widehat{se}\cdot z^*_{\alpha(R+1)}$$

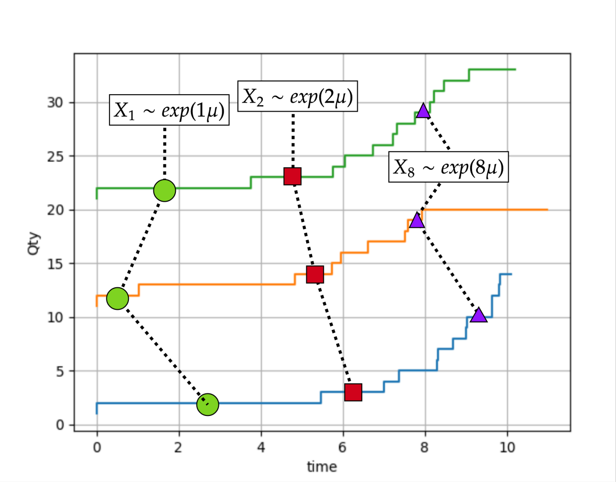
- with an error of $O\left(\frac{1}{nB} + \frac{1}{n^2}\right)$ for $\widehat{se}$ as per earlier.

[2] See [3]

# Increasing reliability

## Goal

If we have $n = 3$ traces, we have at most 3 samples from the same exponential distribution $\rightarrow$ not enough for a reliable bootstrap estimation



# Increasing reliability

## Idea

Convert each independant $X_k \hookrightarrow \exp(k\mu)$ to one single $\tilde{X}_k \hookrightarrow \exp(\mu)$

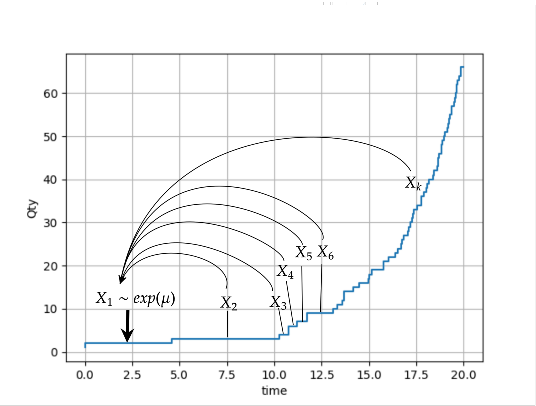



Figure: Petri dish with $k = 1$ bacterium

# Increasing reliability

## Theorem

*If* $X_0 \hookrightarrow \exp(\lambda_0)$, $X_1 \hookrightarrow \exp(\lambda_1)$, $X_0$ *and* $\dfrac{\lambda_1}{\lambda_0} X_1$ *are identically distributed.*

## Proof.

- Let $X_0 \hookrightarrow \exp(\lambda_0)$, $X_1 \hookrightarrow \exp(\lambda_1)$, and $F_{X_0}^{-1}(x) = \int_{-\infty}^{x} \lambda_0 e^{-\lambda_0 t} dt = -\dfrac{1}{\lambda_0} \ln x$ .
- By using the inverse transform sampling relation, we have that $F_{X_0}^{-1}(U) \hookrightarrow \exp(\lambda_0)$ for $U$ any uniformly distributed random variable.
- But $e^{-\lambda_1 X_1} \hookrightarrow \mathcal{U}([0,1])$[3]
- so $F_{X_0}^{-1}\left(1 - e^{-\lambda_1 X_1}\right) = \dfrac{\lambda_1}{\lambda_0} X_1 \hookrightarrow \exp(\lambda_0)$
- So $X_0$ and $\dfrac{\lambda_1}{\lambda_0} X_1$ are identically distributed.

---

[3] See Lemma

# Increasing reliability

$T(X) = [kx_k \ , \ k \in [\![1, n]\!]]$

*Histogram distribution* $e^{-\mu T(X)}$

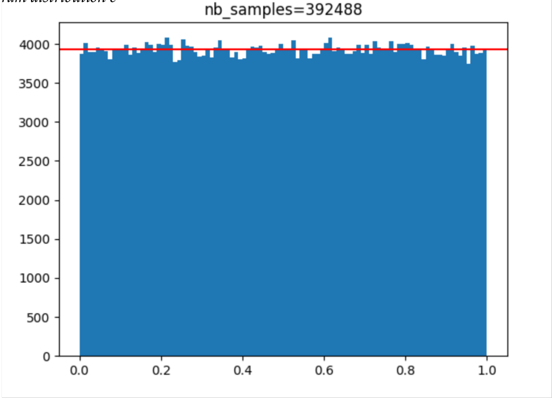



Figure: Petri dish with $k = 1$ bacterium

# Increasing reliability

**Lemma**

$e^{-\lambda_1 X_1} \hookrightarrow \mathcal{U}([0,1])$

# Increasing reliability

- Since each $X_k \sim \exp(k\mu)$, $\frac{k\mu}{\mu} X_k = kX_k$ has the same distribution as $X_1 \sim \exp(\mu)$,
- We have effectively transformed our single trace from 1 exeperiment into *nb_cells* independant samples from $\exp(\mu)$!
- $n = 3 \longrightarrow n = 3 \times nb\_cells \approx 3.10^3$
- $\mathrm{var}(\widehat{se}) < C.10^{-6}$☺☺☺☺☺

# Estimation of $\mu$ on simulations

- Simulations of the boolean network using a Gillepsie-like algorithm:
- 3 Traces/ Experiments, duration =100 secs, $\mu = .5$, Initial Number of cells = 1

# Estimation of $\mu$ on simulations

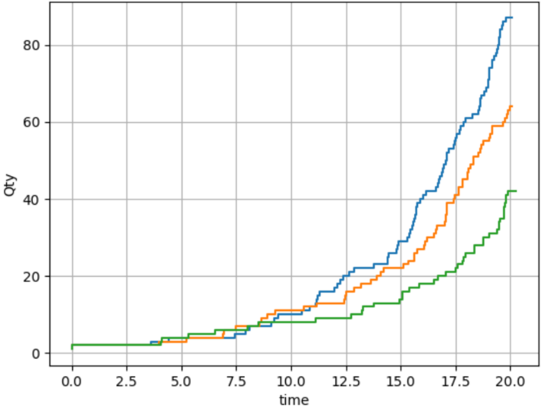



Figure: Petri dish with $k = 1$ bacterium

# Estimation of $\mu$ on simulations

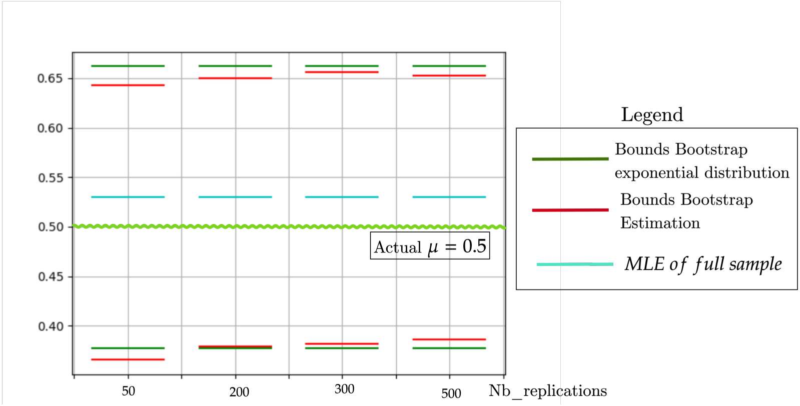



Figure: Petri dish with $k = 1$ bacterium

# Discussions

1. Need to think over the assumption of "representativeness" of the empirical distribution
2. Need to be adapted to reality : Interpolation
3. We worked on a simple model : $A \xrightarrow{\mu} 2A$. More complex ones? Resource consumption model?

- Stochastic Modelling for Systems Biology Third Edition Darren J. Wilkinson
- Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC
- 'Bootstrap Methods and their Application', by A. C. Davison and D. V. Hinkle Cambridge University Press, 1997
- Stochastic Rate Parameter Inference Using the Cross-Entropy Method, Jeremy Revell and Paolo Zuliani