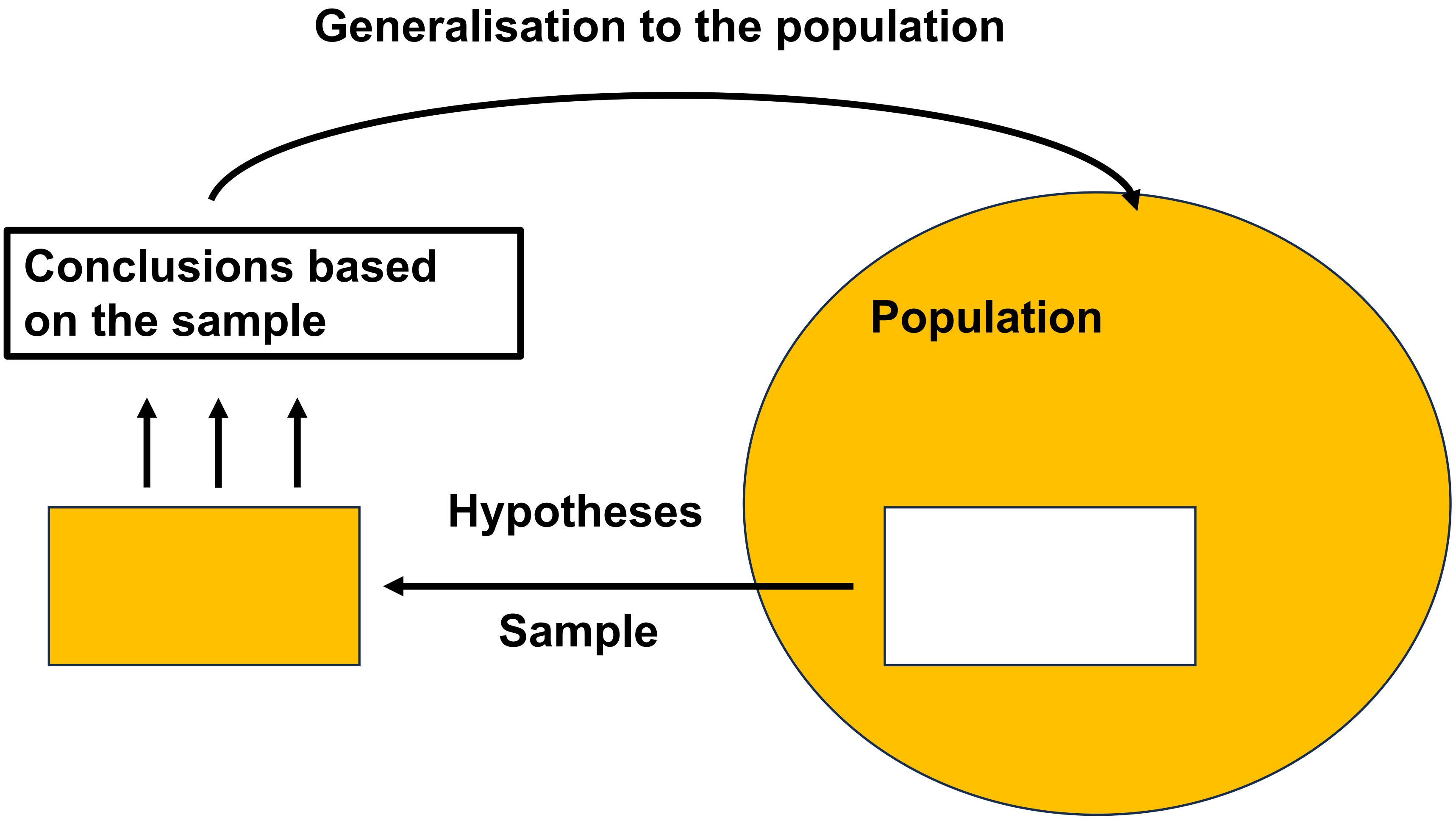
Random variable, maximum Likelihood

Nicolas Rode

JC CBGP, 27 Mar. 2024

Statistical inference



Generalisation to the population

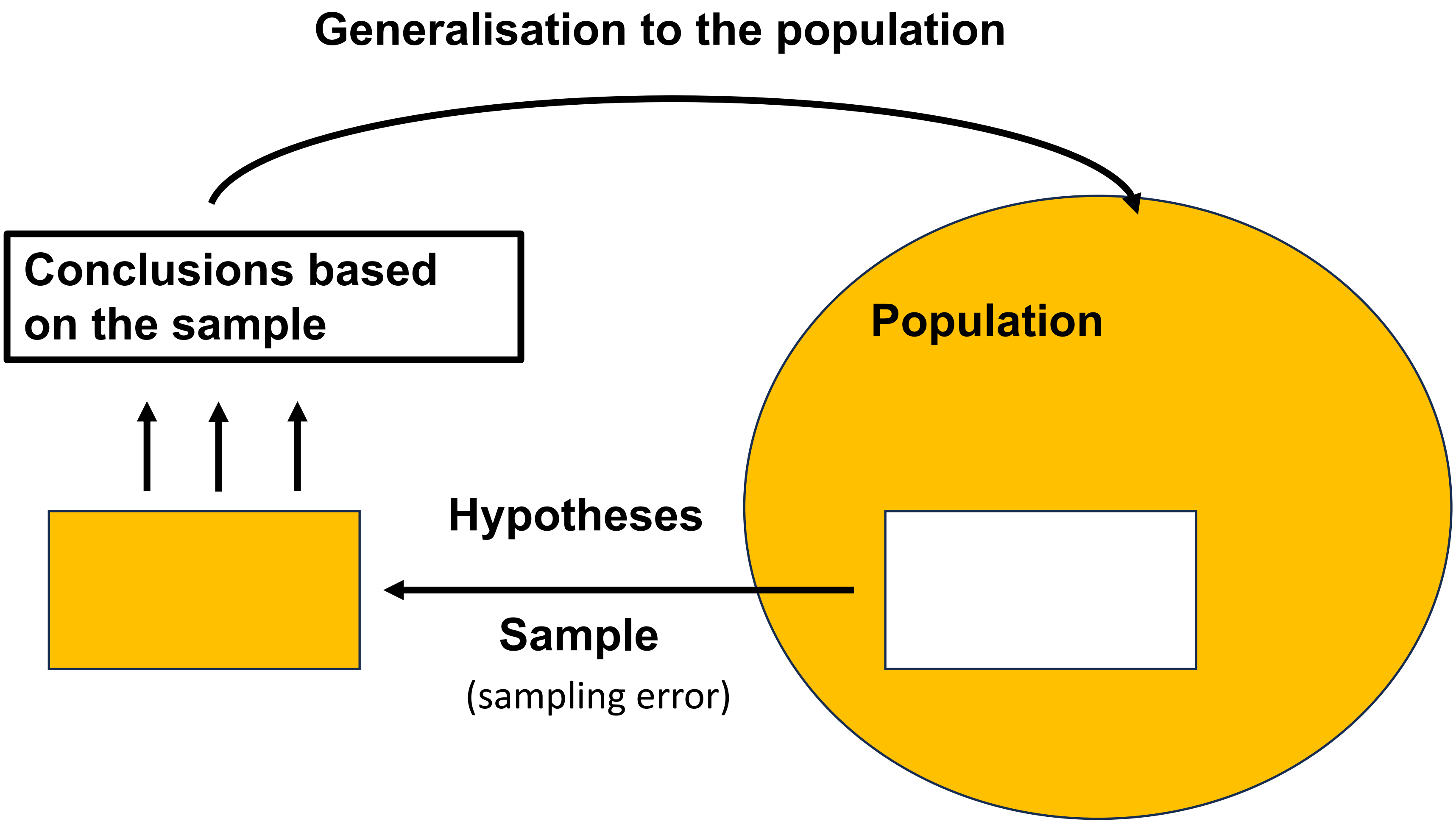
Conclusions based on the sample

Population

Hypotheses

Sample

Statistical inference



Generalisation to the population

Conclusions based on the sample

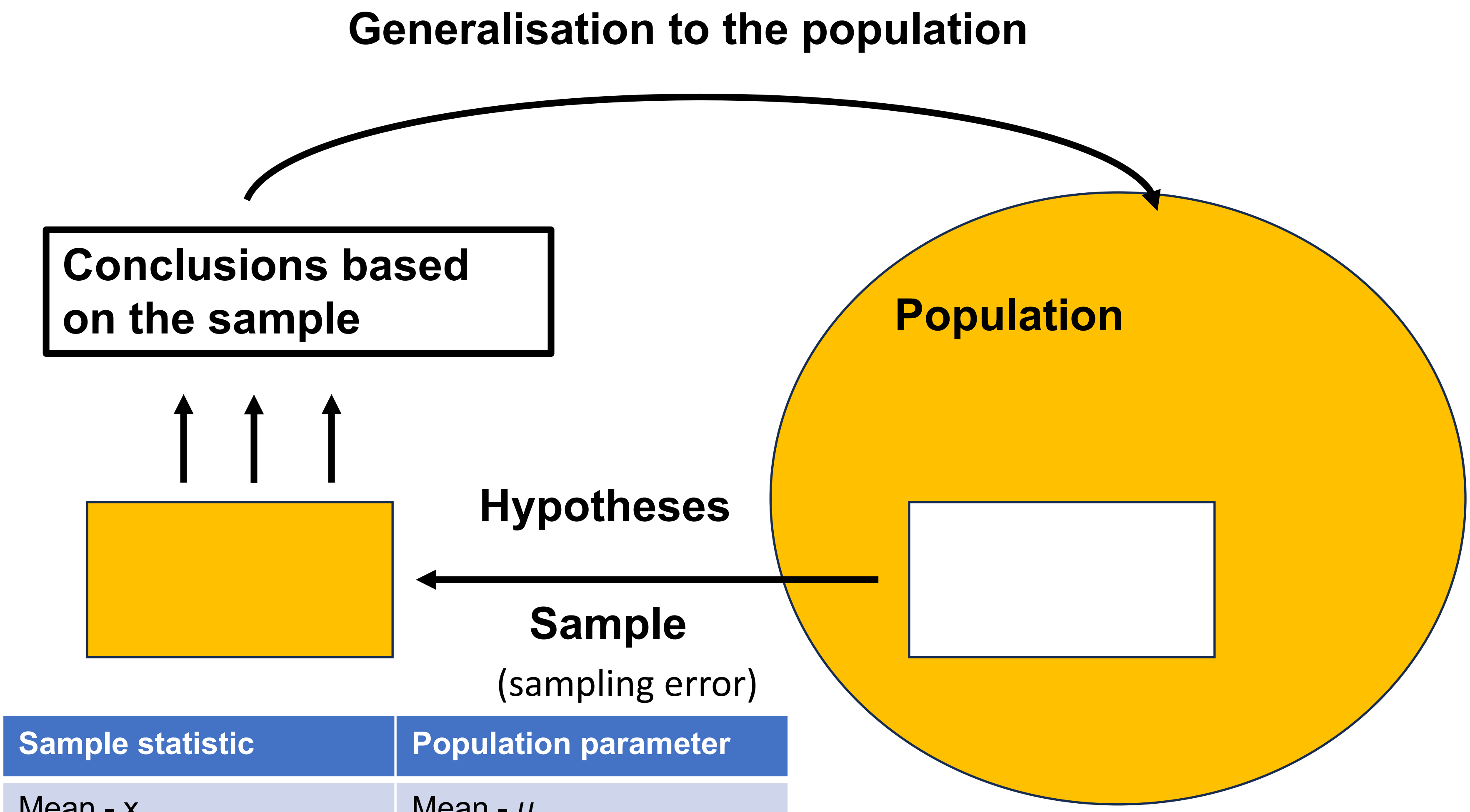
Population

Hypotheses

Sample

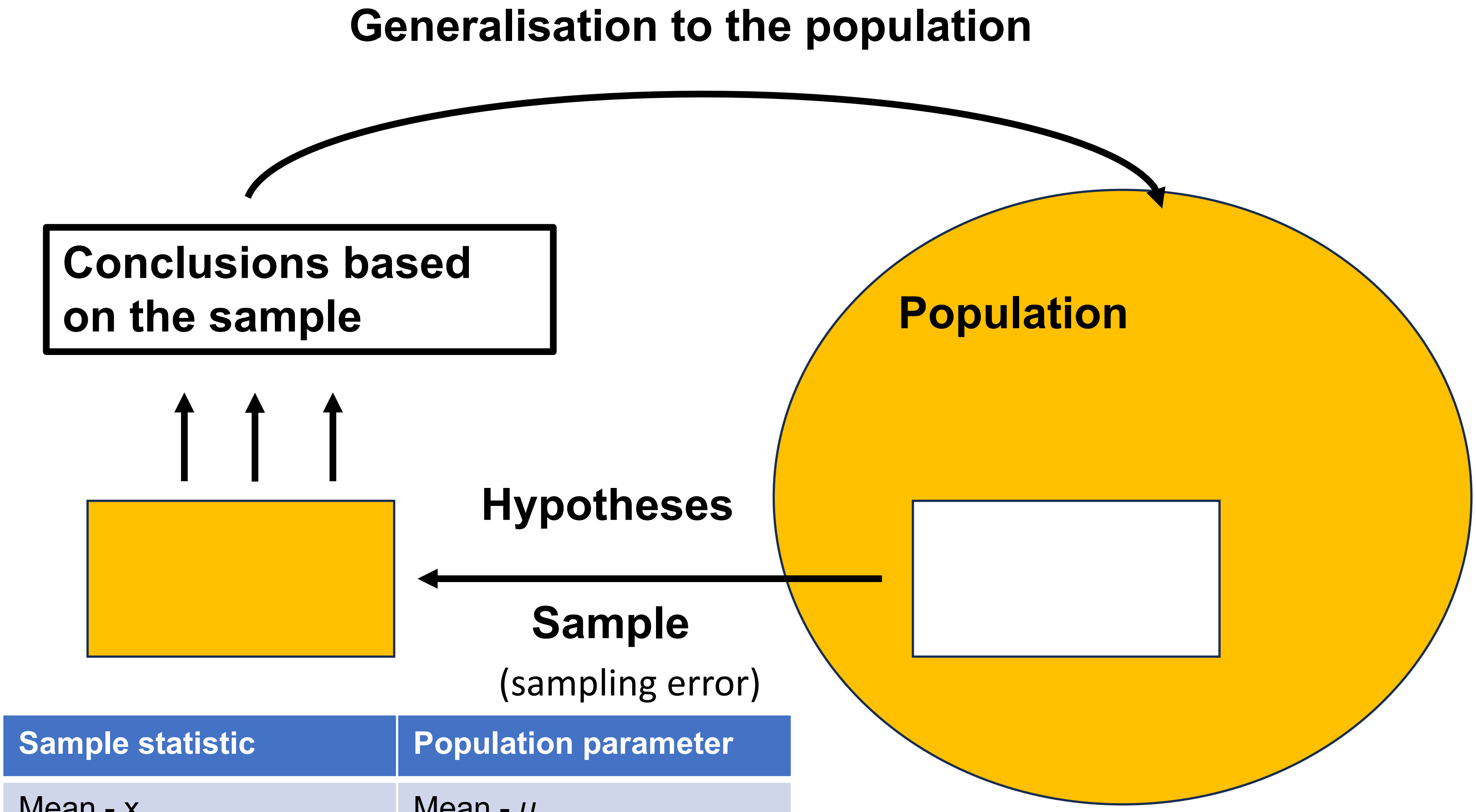
(sampling error)

Statistical inference


Generalisation to the population
Conclusions based on the sample
Population
Hypotheses
Sample
(sampling error)

Sample statistic	Population parameter
Mean - x	Mean - μ
Variance - s^2	Variance - σ^2
Standard deviation - s	Standard deviation - σ

Statistical inference


Generalisation to the population
Conclusions based on the sample
Population
Hypotheses
Sample
(sampling error)

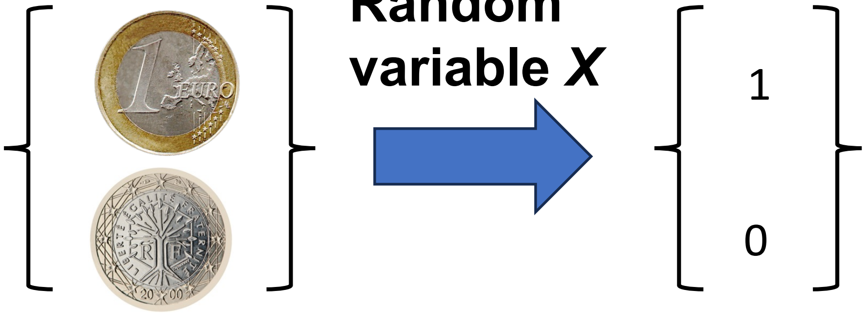
Sample statistic	Population parameter
Mean - x	Mean - μ
Variance - s^2	Variance - σ^2
Standard deviation - s	Standard deviation - σ

Point estimate=$\hat{\mu}$

Interval estimate =$\hat{\mu} \pm 1.96 \times SE$

What is a random variable *X*?

Definition: function which depends on the outcome of a random process during an experiment. It associates a **number with each outcome of the random process.**



Random variable *X*

1

0

$$\mathrm{P}[x_i = 1] = \frac{1}{2}$$

$$\mathrm{P}[x_i = 0] = \frac{1}{2}$$

Evènement dans Ω,
l'univers des possibles

Valeur réelle x_i

Probability

$$x_i \in \{0; 1\}$$

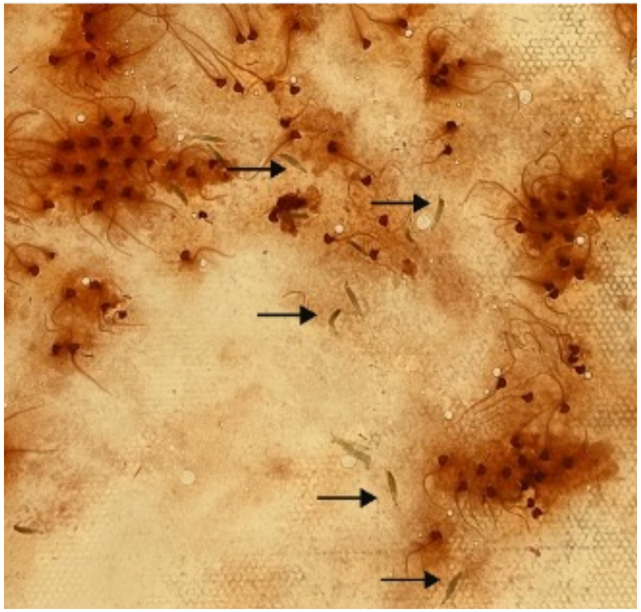
What is a random variable X?

Number of larvae in a strawberry

Zriki et al 2023 *J. Econ. Entom.*

Evènement dans Ω,
l'univers des possibles

Random variable X



Valeur réelle x_i

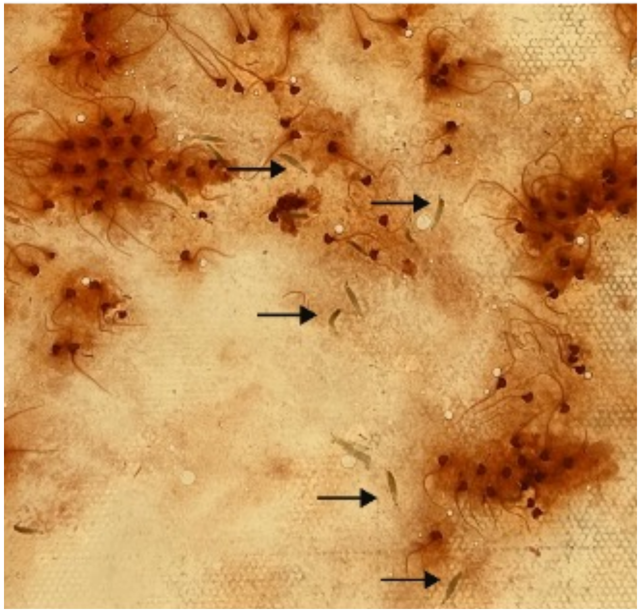
=> Random variable X with discrete values $x_i \in \{0, 1, 2, \ldots\}$

=> Probability distribution of X (=its probability law) $(x_i, \mathrm{P}[X = x_i])$

Probability distribution for count data

Poisson distribution

Zriki et al 2023 *J. Econ. Entom.*



Number of larvae in a strawberry

Models count data where the events occur randomly and independently at a constant rate

$$x_i \in \{0, 1, 2, \dots\}$$

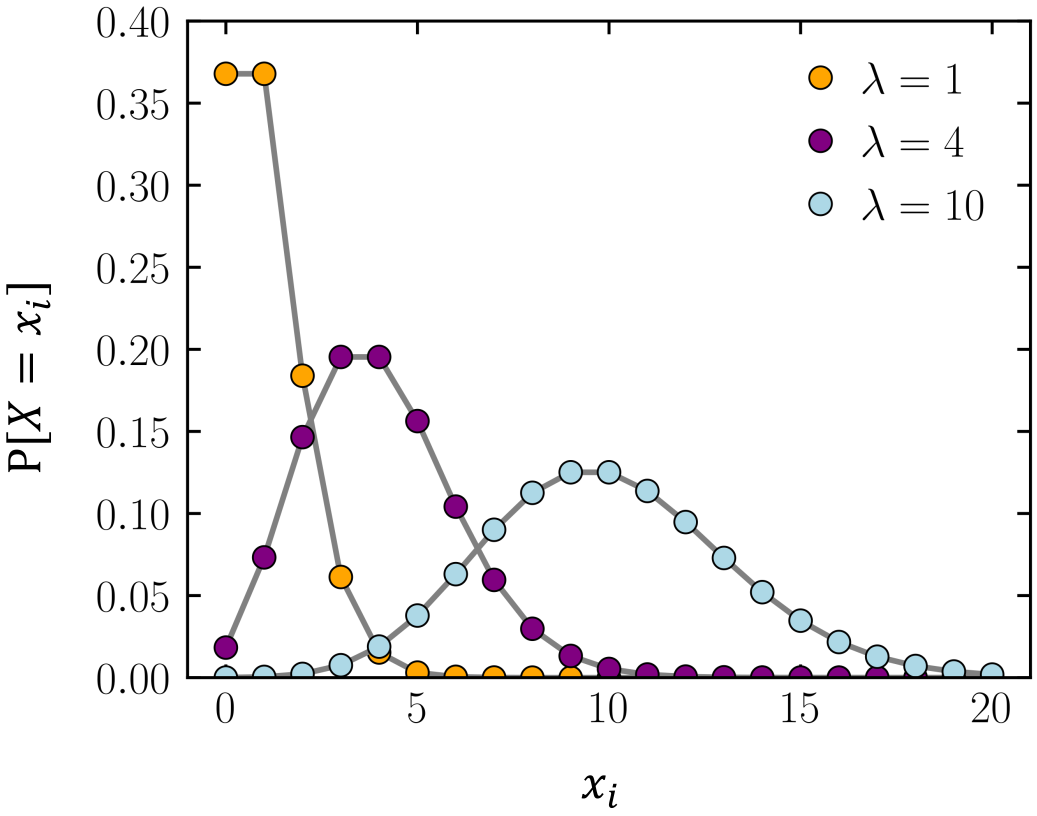
Examples: Number of larvae, number of parasites, number of species

Probability law function: $\mathrm{P}[X = x_i] = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

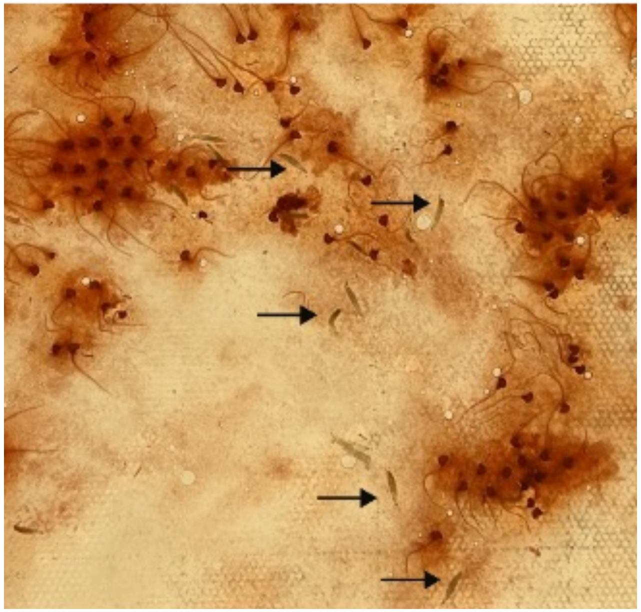
Assumptions: The mean and variance of the distribution are equal

$$\mathrm{E}[X] = \mathrm{V}[X] = \lambda$$

What is the probability distribution function of a discrete random variable X?



Zriki et al 2023 *J. Econ. Entom.*



Valeur réelle x_i

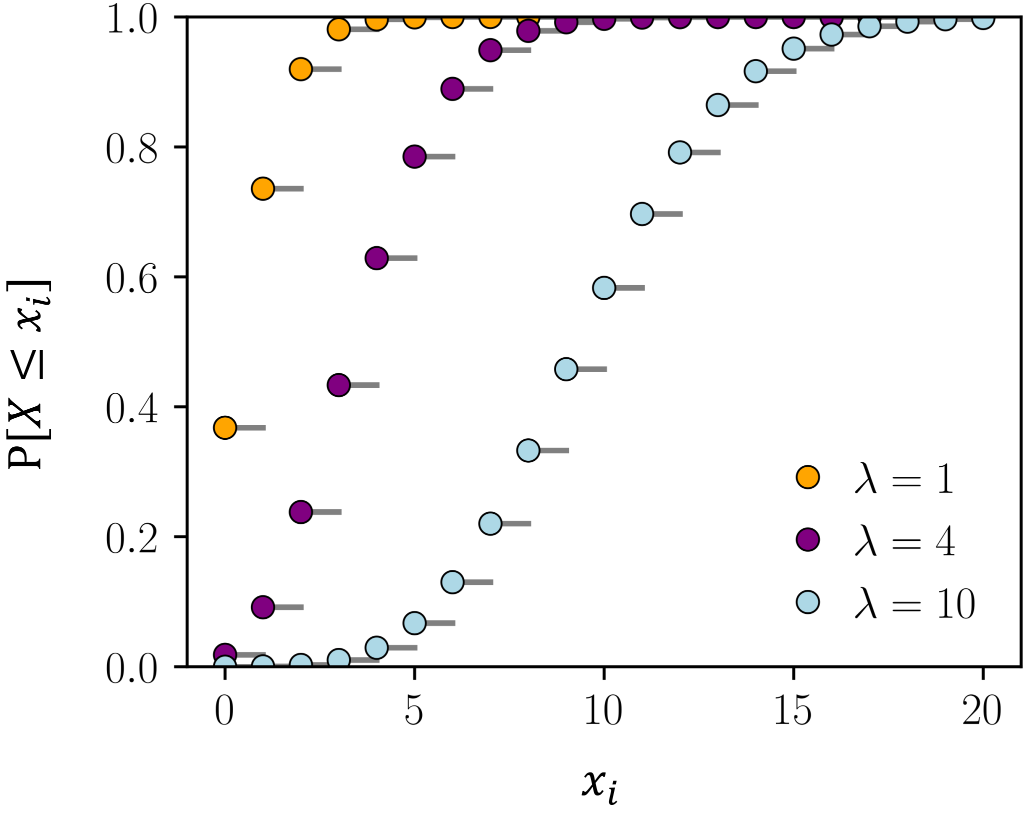
Probability law function: $\mathrm{P}[X = x_i] = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

$x_i \in \{0, 1, 2, \ldots\}$

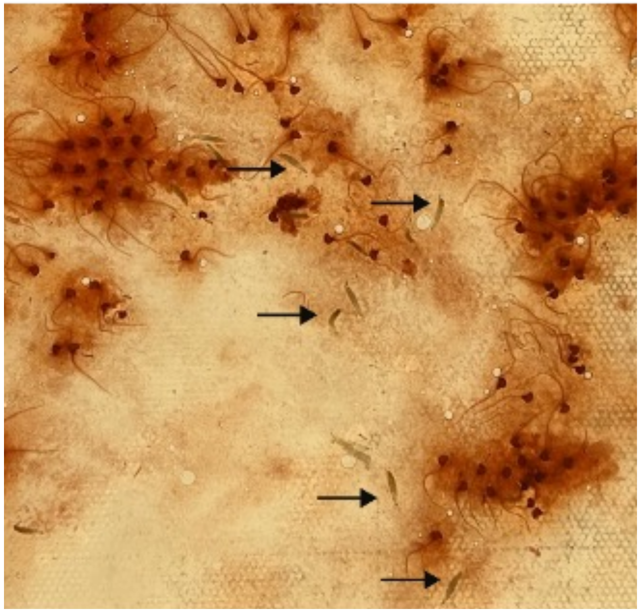
Assumptions: The mean and variance of the distribution are equal

$$\mathrm{E}[X] = \mathrm{V}[X] = \lambda$$

What is the cumulative distribution function of a random variable X?



Zriki et al 2023 *J. Econ. Entom.*



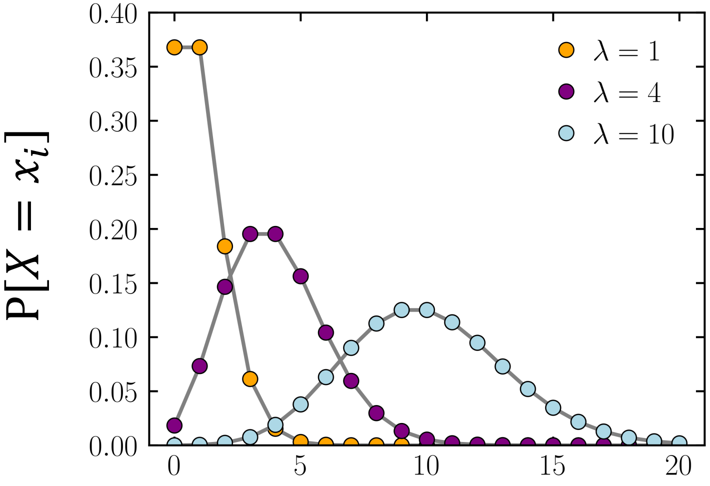
Valeur réelle x_i

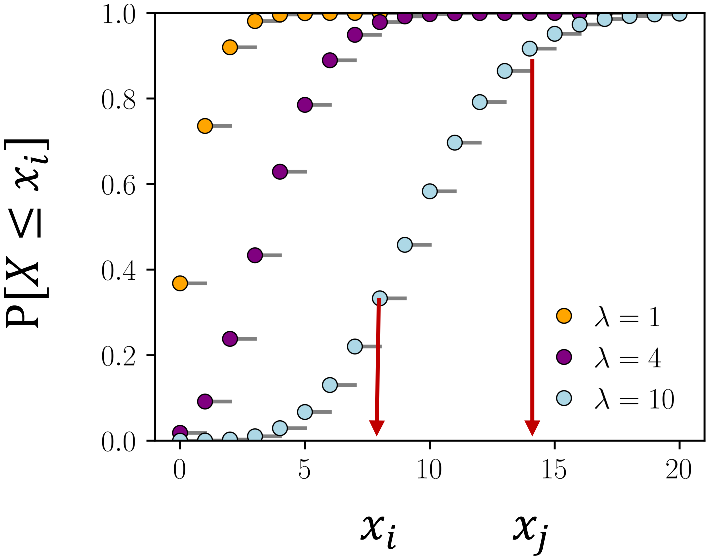
$x_i \in \{0, 1, 2, \dots\}$

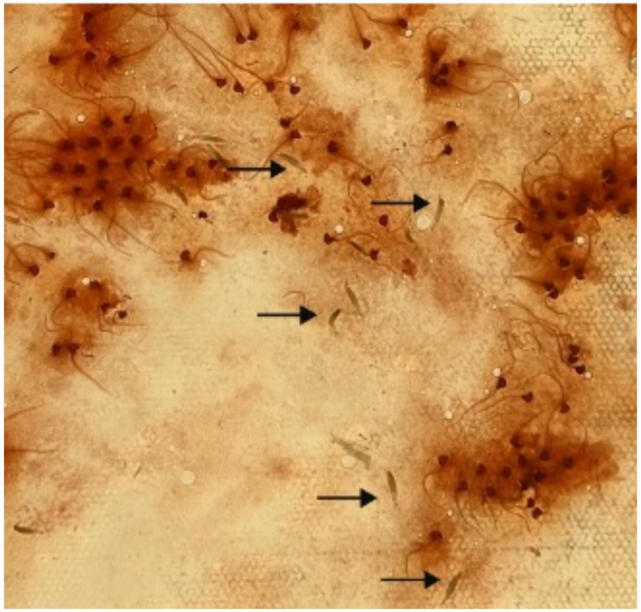
Definition: probability that the random varibale X will take a value less than or equal to x

What is the difference between the cumulative and probability distribution functions of X?

Zriki et al 2023 *J. Econ. Entom.*







Valeur réelle x_i

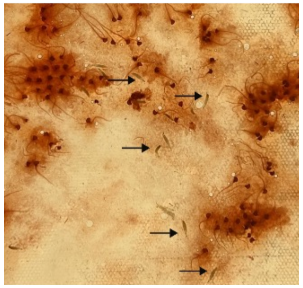
$x_i \in \{0, 1, 2, \ldots\}$

If $x_i < x_j$, $\mathrm{P}[x_i < X \leq x_j]$= $\mathrm{P}\left[X \leq x_j\right]$- $\mathrm{P}[X \leq x_i]$

$\mathrm{P}\left[X \leq x_j\right] = \mathrm{P}[X \leq x_i]$+$\mathrm{P}[x_i < X \leq x_j]$

Travaux pratiques!

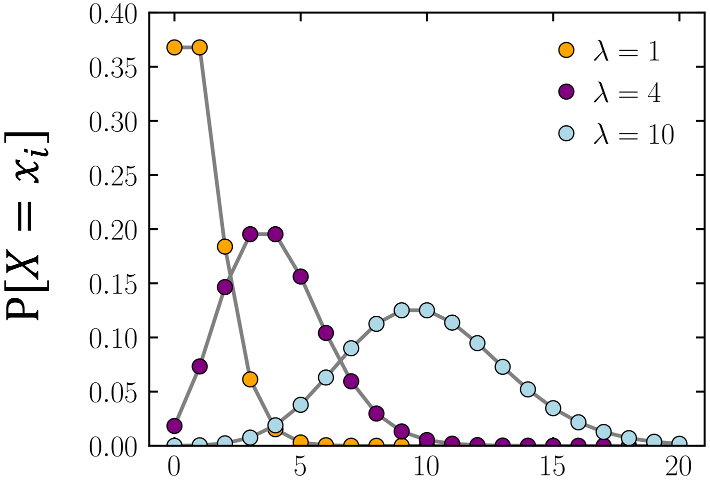
Q1: A research scientist at CBGP studies the rate of infestation of fruits by an insect pest. He assumes that each fruit is infested by one larva on average. What is the probability of observing zero larva in one fruit? One larva? Two larvae?



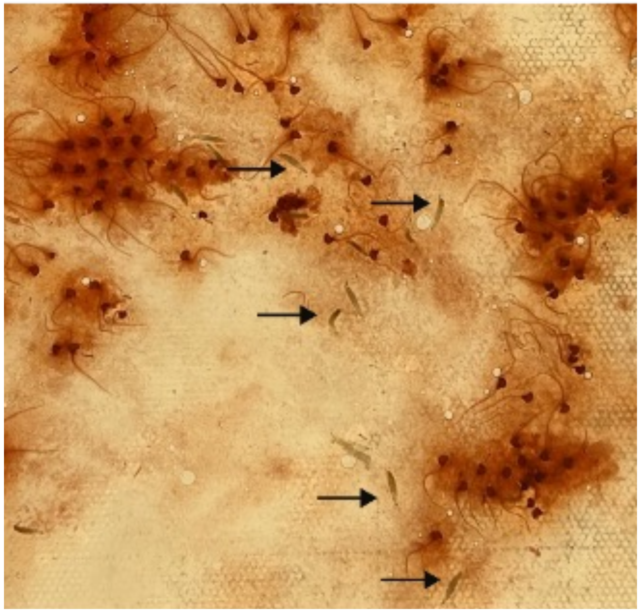
Q2: A PhD student at CBGP wants to know whether its fruit purée is pasteurized. She plates 10uL of purée on one Petri dish and observes no microbial colony. What is the 95% confidence interval of the average number of colonies she can expect in 10uL of purée?

Q3: A PhD student at CBGP wants to estimate the rate of infestation by a helminth parasite in a population of rodents. He found no parasite among 10 individuals screened. What is the approximate 95% confidence interval of the rate of infestation in this population? What would be the 95% CI if he would have found 10 infected individuals?

What is Maximum Likelihood Estimation?



Zriki et al 2023 *J. Econ. Entom.*



Definition: Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters of a statistical model by finding the parameter values that maximize the likelihood function, given the observed data. In other words, MLE seeks to find the parameter value, $\bar{\lambda}$, that makes the observed data most probable under the assumed statistical model.

Poisson example

- Suppose that we know that the following five numbers were simulated using a Poisson distribution: 0 2 1 2 3
  We can denote them by y1, y2, ..., y5. So y1 = 0 and y5 = 3
- Recall that the pdf of a Poisson random variable is
- f(x_i ; λ) = $\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$, where $x_i \in \{0, 1, 2, \dots\}$
  For example, f(0 ; λ) =$e^{-\lambda}$
- We want to figure out what is the λ that was used to simulate the
- five numbers
- All we know is that
  1) they come from a Poisson distribution and
  2)they are independent from each other
- Since we know the pdf that generated the numbers is Poisson, we know that the probability of the first number is $\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}$,
- The probability of the second is $\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}$ and so on...
- We could replace the x_i with the actual numbers. For example, the first one is y1 = 0 so the probability is just $e^{-\lambda}$. We will keep the symbols because we are going to make the problem more general
- What we do not know is the value of the parameter λ

Poisson example

- Since we know that they are independent we could also write down the probability of observing all 5 numbers. That is, their joint probability.
- Since they are independent their joint distribution is the multiplication of the 5 pdfs. Recall: p(A ∩ B) = P(A)P(B) if A and B are independent.
- We use the product symbol ∏ to simplify the notation. For example, $\prod_{i=1}^{i=n} x_i = x_1 \times x_1$
- So we can write the joint probability or the likelihood (L) of seeing those 5 numbers as:

$$\mathrm{L}(\lambda) = \prod_{i=1}^{i=5} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

- Remember that we are trying to find λ that was used to generate the 5 numbers. That's our unknown.
- In other words, we want to find the λ that maximizes the likelihood function $\mathrm{L}(\lambda)$. Once we find it, we could it write our estimated parameter as $\bar{\lambda}$
- Yet another way: we want to find the $\bar{\lambda}$ that makes the joint likelihood of seeing those numbers as high as possible
- Sounds like a calculus problem... We can take the derivative of L(p) with respect to λ and set it to zero to find the optimal $\bar{\lambda}$
- Of course, the second step is to verify that it's a maximum and not a minimum (take second derivative) and also verify that is unique, etc.
  We will skip those steps.

Poisson example

- Taking that derivative is complicated because we would need to use the chain rule several times. A lot easier to make it a sum so we take the log; the log function is a monotonic transformation, it won't change the optimal $\bar{\lambda}$ value
- We will use several properties of the ln, in particular:

$$\ln(x^a y^b) = \ln(x^a) + \ln(y^b) = a\ln(x) + b\ln(y)$$

- So now we have (for n numbers rather than 5):

$$\text{ln}\mathrm{L}(\lambda) = \sum_{i=1}^{i=n} -\lambda\ln(e) + x_i\ln(\lambda) - \ln(x_i!)$$

- Which simplifies to: $\text{ln}\mathrm{L}(\lambda) = -n\lambda + n\bar{x}\ln(\lambda) - \sum_{i=1}^{i=n}\ln(x_i!)$
- This looks a lot easier, all we have to do is take $\frac{d\text{ln}\mathrm{L}(\lambda)}{d\lambda}$,

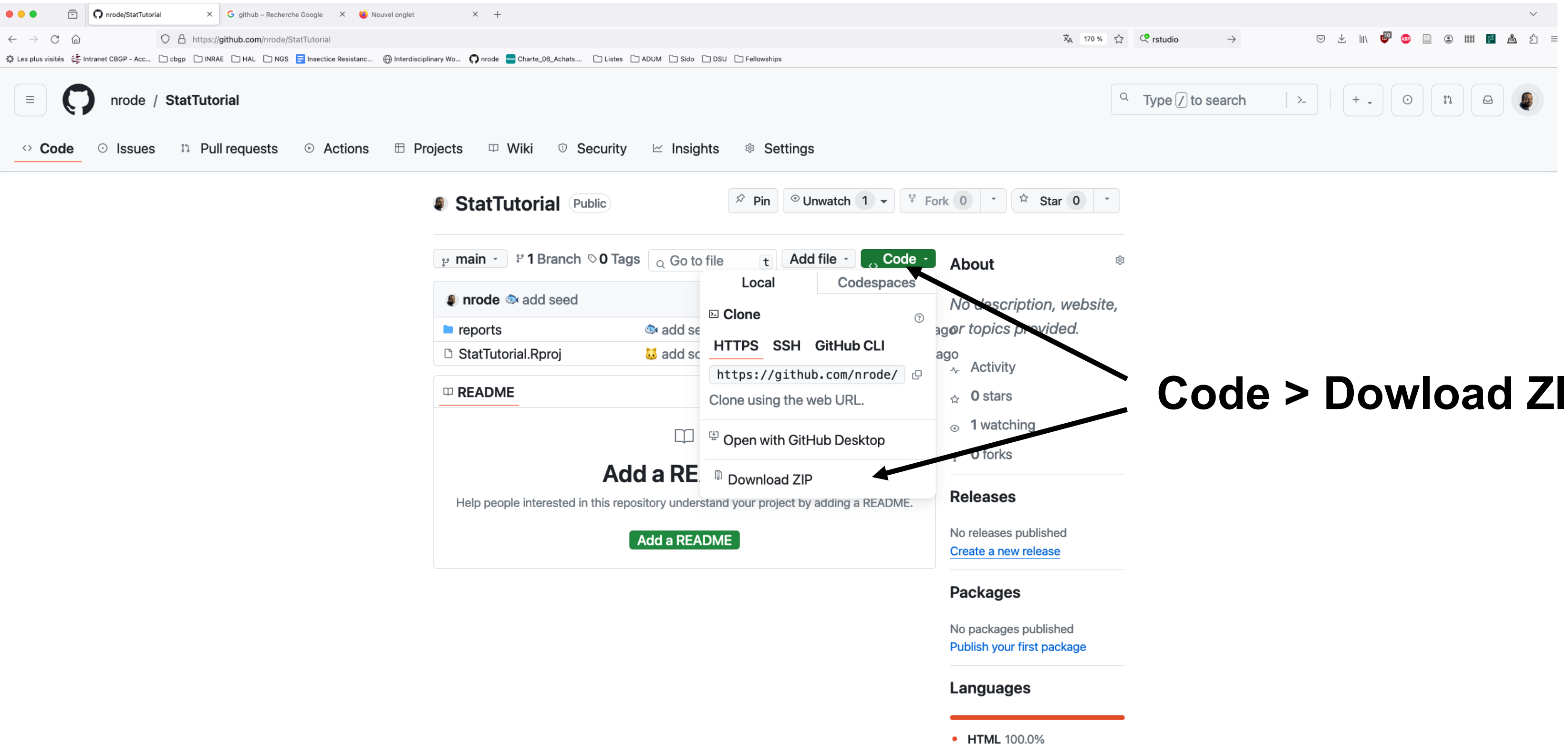
set it to zero, and solve for λ

$$\frac{d\text{ln}\mathrm{L}(\lambda)}{d\lambda} = -n + \frac{n\bar{x}}{\lambda} = 0$$

- After solving, we'll find that $\bar{\lambda} = \bar{x} = \sum_{i=1}^{i=n} x_i$
- So that's the MLE estimator of λ. This is saying more or less the obvious: our best guess for the λ that generated the data is the average number of counts, in this case $\bar{\lambda} = 1.6$
- Note that we can plug in the optimal $\bar{\lambda}$ back into the log-likelihood function: $\text{ln}\mathrm{L}(\bar{\lambda}) = -n\bar{\lambda} + n\bar{x}\ln(\bar{\lambda}) - \sum_{i=1}^{i=n}\ln(x_i!) = a$, where a will be a number that represents the highest likelihood we can achieve (we found $\bar{\lambda}$ that way)

Travaux pratiques!

https://github.com/nrode/StatTutorial



Code > Dowload ZIP

```
https://github.com/nrode/
```

Clone using the web URL.

Add a README

Help people interested in this repository understand your project by adding a README.

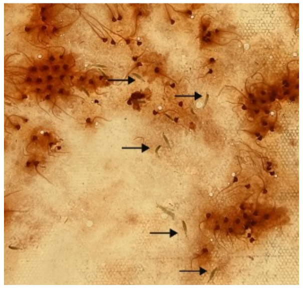
No description, website, or topics provided.

No releases published

No packages published

Travaux pratiques!

Q4: Simulate infestation with an average of one larva per fruit and estimate the parameter used for the simulations using maximum likelihood



References

Peraillon MC (2020). Health Services Research Methods I. Lecture notes. https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press. p163 https://math.mcmaster.ca/~bolker/emdbook/index.html

Lessells, C. M., & Boag, P. T. (1987). Unrepeatable repeatabilities: a common mistake. *The Auk*, *104*(1), 116-121. https://www.researchgate.net/profile/Peter-Boag-2/publication/231424296_Unrepeatable_Repeatabilities_A_Common_Mistake/links/0912f5069dcf9a594b000000/Unrepeatable-Repeatabilities-A-Common-Mistake.pdf

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935-956. https://doi.org/10.1111/j.1469-185X.2010.00141.x

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution*, *3*(1), 129-137. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210X.2011.00125.x

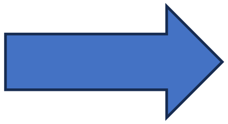
Zriki, G., Blois, R., Fournier, C., Tregoat-Bertrand, L., Poupart, P. Y., Bardel, A., ... & Rode, N. O. (2023). A fast and reliable larval sampling method for improving the monitoring of fruit flies in soft and stone fruits, *Journal of Economic Entomology*, 2024;, toae001, https://doi.org/10.1093/jee/toae001

Probability distribution for proportion data

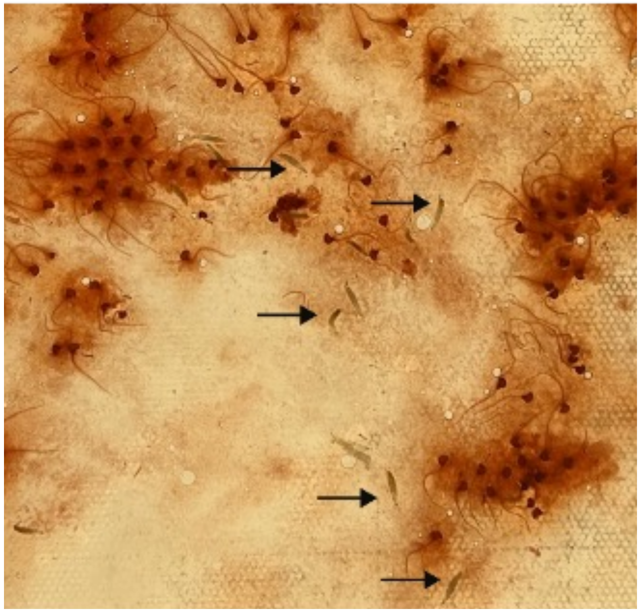
Binomial distribution

Zriki et al 2023 *J. Econ. Entom.*

Evènement dans Ω,
l'univers des possibles



Random variable *X*



Valeur réelle x_i
(number of infested strawberries)

$x_i \in [0, \dots n]$

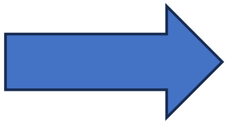
=> Random variable *X* with discrete values

Probability distribution for Gaussian data

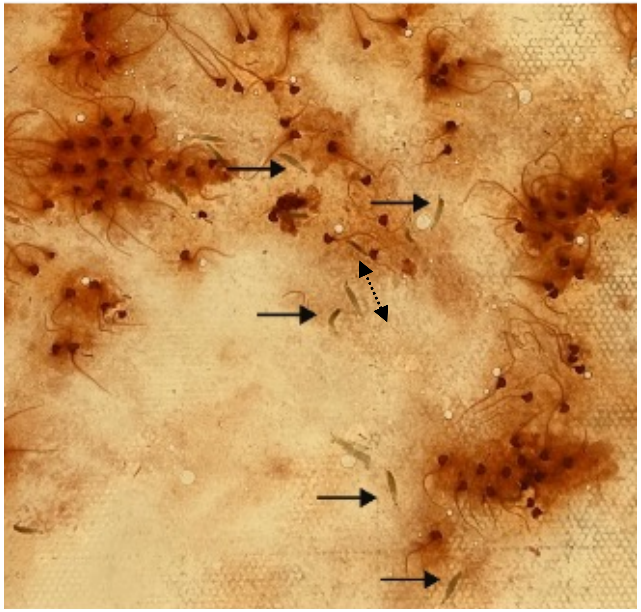
Binomial distribution

Zriki et al 2023 *J. Econ. Entom.*

Evènement dans Ω,
l'univers des possibles



Random variable *X*



Valeur réelle x_i
(size of one larva)

$x_i \in \mathbb{R}$

=> Random variable *X* with continuous values