Day 4
Lecture 2:
Using outbreak data

Short course on modelling infectious disease dynamics in R

Ankara, Türkiye, September 2025

Dr Juan F Vesga

Aims of the session

- Understand what attributes outbreak data has and how it should be analysed
- Familiarise with important delays like incubation period, generation time and serial interval
- Understand some of the statistical details behind building this indicators
- Understand Case fatality Ratios
- Understand how R0 can be estimated from Epicurves

Early outbreak context

- within a few days / weeks of index case
- limited data available
- no or limited intervention
- no depletion of susceptibles
- urgent assessment needed to inform response

Key questions

- Disease-dependent, but generally includes:
  - How fast is it growing?
  - What is driving the epidemic growth?
  - What is the case fatality ratio?
  - Who is most severely affected?
  - How many cases should we expect in the next days / weeks?

Some basic definitions

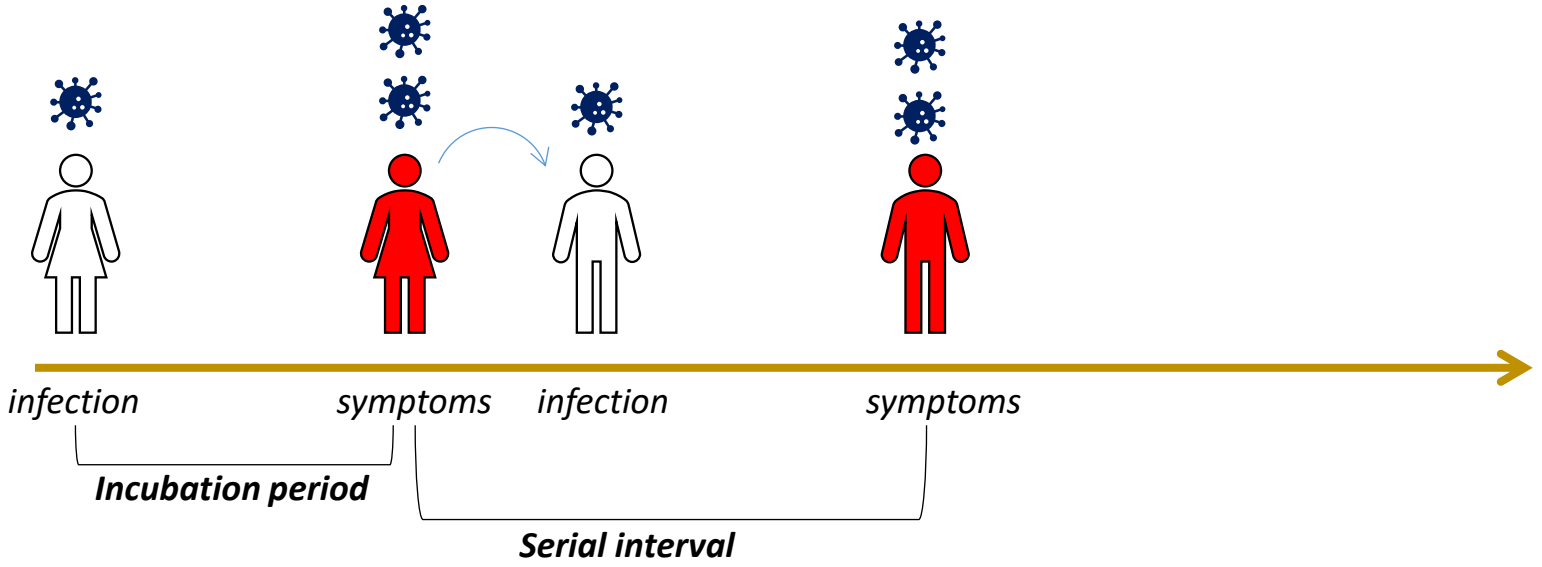
- **population**: set of all possible observations of a given process/entity
  example: all possible cases of cholera in location xxx
- **sample**: subset of the population
  example: all cases of cholera in xxx reported last week
- a **statistic**: quantity used to describe sample / population
  example: % of fatalities in cholera cases in xxx last week
- **inference**: statement about population(s) from sample(s)
  example: % of fatalities in cholera cases in xxx is greater than in yyy

The incubation period

- Definition: time interval between the date on infection and the date of symptom onset

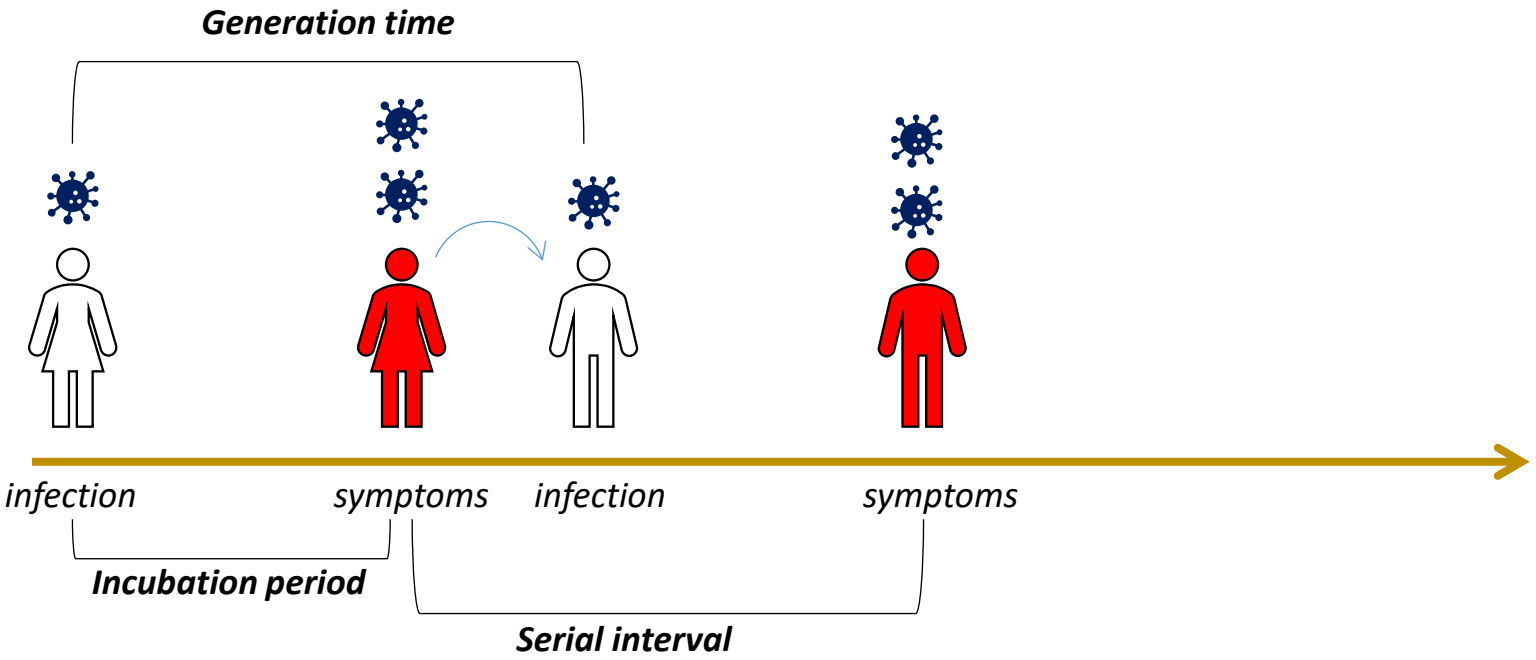
The serial interval

- Definition: time interval between onset of symptoms in primary and secondary cases

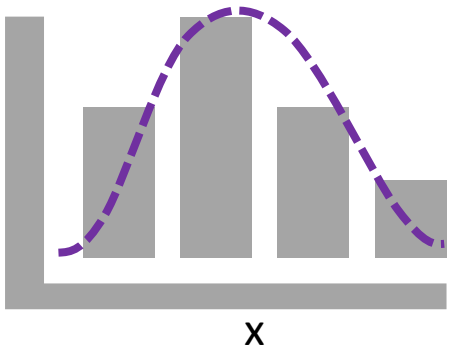


The generation time

- Definition: time interval between date of infections in primary and secondary cases

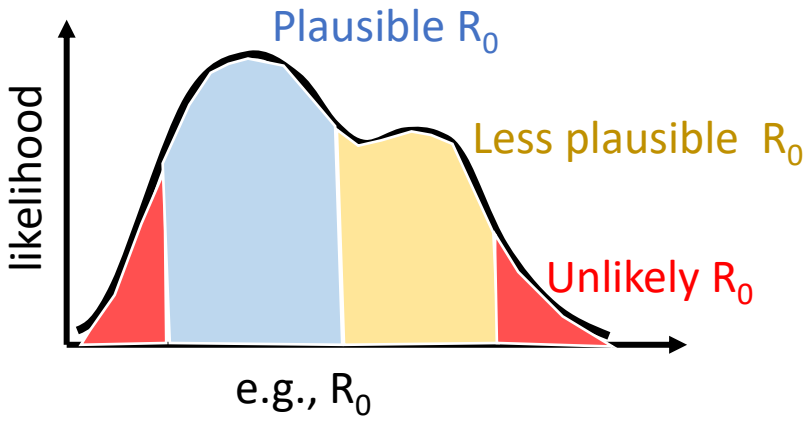


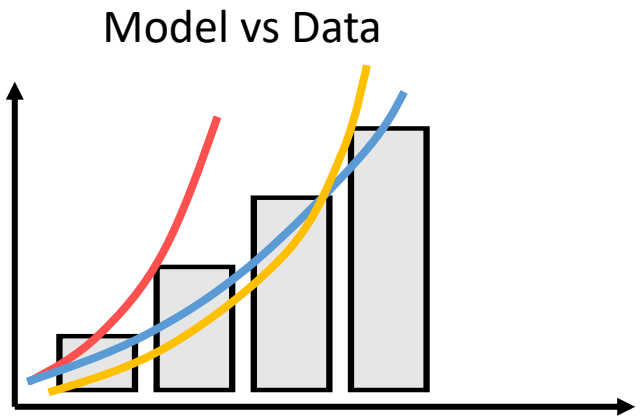
Estimating the underlying distribution



- choose type of distribution (e.g. normal, Poisson, Gamma)
- find θ_x which maximise *p(x)*, i.e. the likelihood
- visually: best fit between bars (data) and curve (distribution)

What is a likelihood?

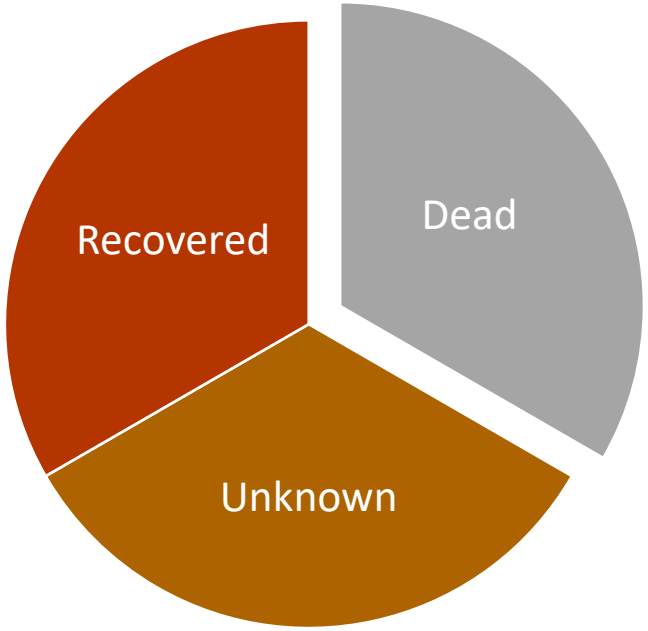




- Likelihood: A relative measure of fit between data and model
- *L = p(Data|Model)*

Case fatality ratio

- **Definition:** the proportion of cases who die of the infection.



$$CFR = \frac{D}{R + D}$$

Case fatality ratio - caveats

- "*case fatality rate*": this is a proportion, not a rate
- Wrong denominator underestimates CFR

Right

$$CFR = \frac{D}{R + D}$$

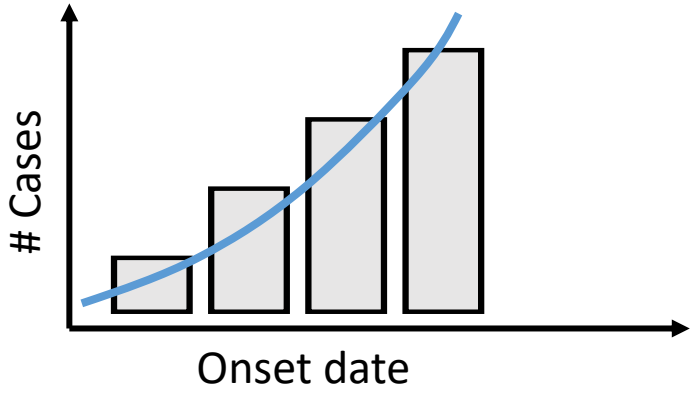
Wrong

$$CFR = \frac{D}{R + D + U}$$

- not accounting for uncertainty, e.g. comparing CFR across groups without statistical tests

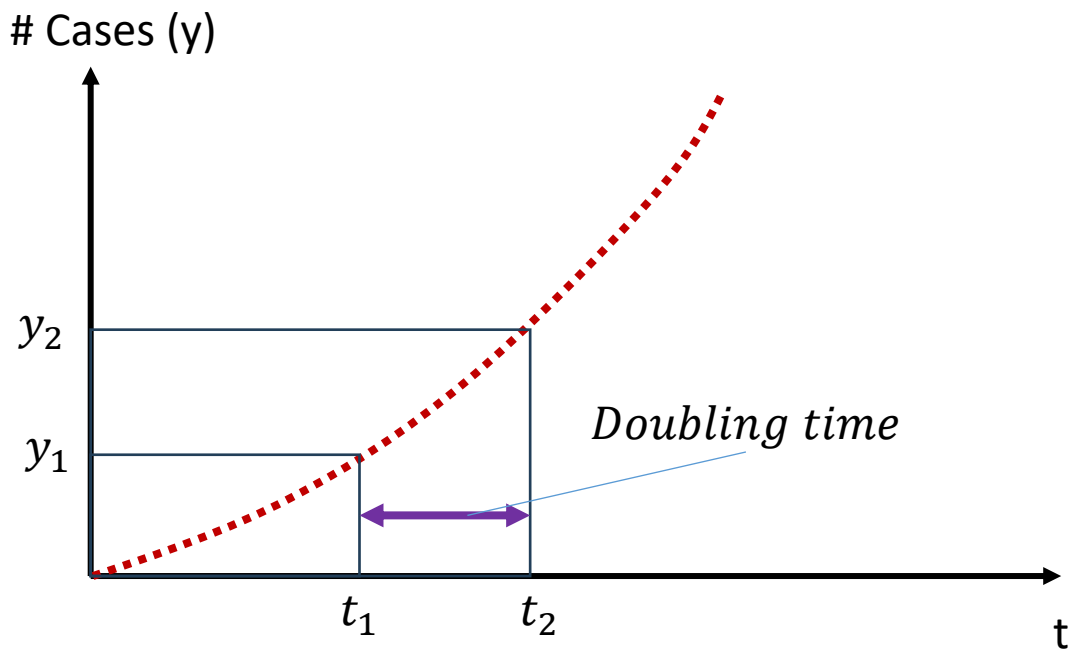
Analysing Epicurves (incidence)

- **Definition**: the incidence is the number of new cases on a given time period.
- relies on dates, typically of onset of symptoms
- only daily incidence is non-ambiguous
- other definitions (e.g. weekly) rely on a starting date
- **prone to reporting delays**



Doubling time

- Let T be the time taken by the incidence to double, given a daily growth rate r



$$\frac{y_2}{y_1} = 2 \Leftrightarrow$$

$$\frac{e^{rt_2+b}}{e^{rt_1+b}} = 2 \Leftrightarrow$$

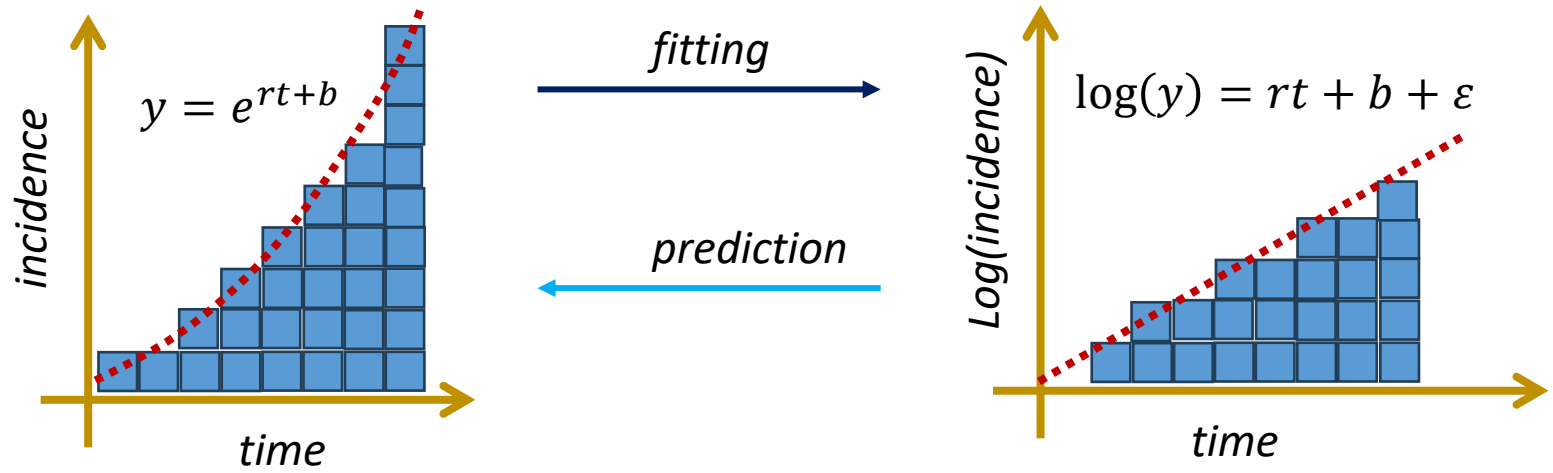
$$e^{rt_1+b} = 2 \Leftrightarrow$$

$$e^{r(t_2-t_1)} = 2 \Leftrightarrow$$

$$T = \log(2)/r$$

Log-linear model of incidence

- Handy form for fitting to incidence curve data



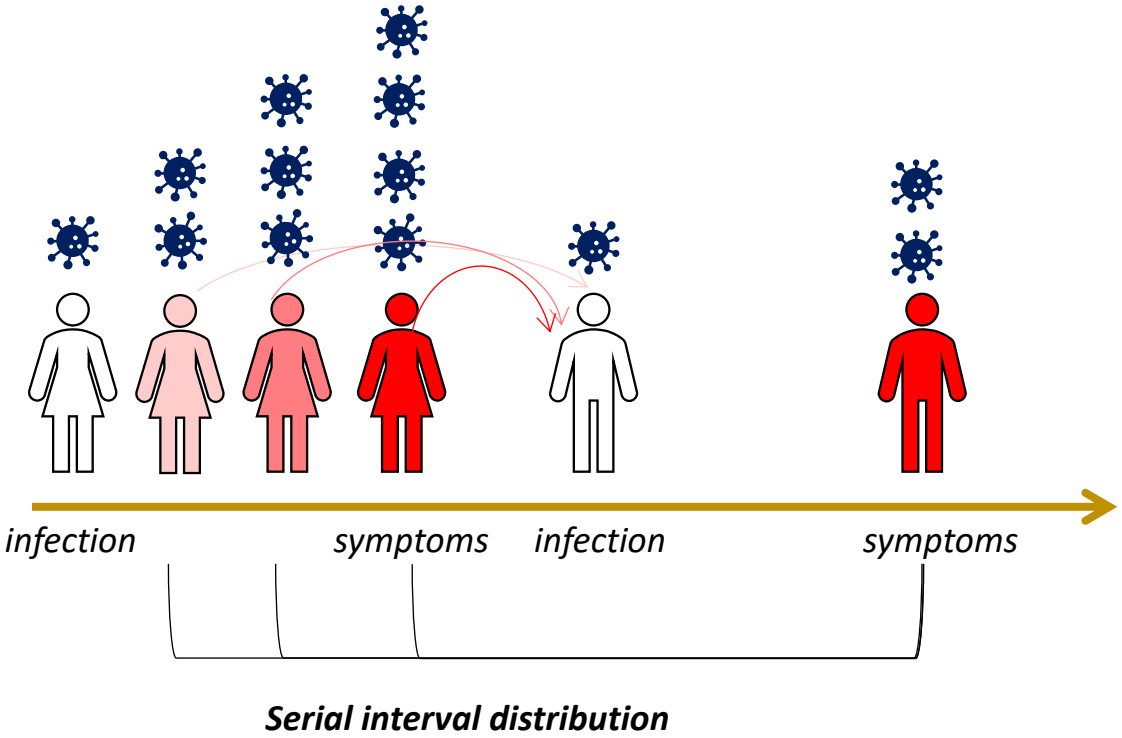
r: $growth\ rate$
b: $intercept$
$\varepsilon \sim N(0, \sigma_\varepsilon)$

Log-linear model: pros and cons

- Pros:
  - fast and simple
  - predictions possible
  - doubling / halving time readily available
  - possible extensions to estimate R0 from r
- Cons:
  - zero incidence problematic
  - non mechanistic
  - no inclusion of other information (e.g. serial interval)

Global infectiousness over time

- Distribution of serial interval has a impact on global infectiousness over time



$$\lambda_t = R_0 \sum w(t - t_i)$$

λt: global force of infection;
w(): serial interval distribution;
ti: date of symptom onset

Summary

- Epicurves summarize and track the evolving outbreak and are the base for most important estimates in outbreak analysis
- A systematic collection of data in linelists provide the basis for key delays like serial interval, incubation period and generation time
- Similarly re cording death outcomes allow us to estimate CFR
- Analysis of incidence curves using the log linear model is the starting point to estimating R0 from outbreak data