L3, Inference on stochastic epidemic models

Tom Britton

July, 2022

Statistical inference/estimation in general

Stochastic modelling can tell us (within a model and given some parameter values): what are the likely outcomes?

Example: Given R_0, about how many will get infected?

Statistical inference goes in the "opposite direction" (within a certain model): given an observed outcome, which parameter "fits" to the observation best?

Example: Suppose 20% were infected during an outbreak. What is R_0?

Estimation from outbreak sizes

Suppose an epidemic outbreak is observed and we want to estimate parameters, e.g. transmission probability p, or R_0

What is observed?

Final size: how many were infected and how many were not during outbreak

Important with additional knowledge of how many/what fraction were susceptible prior to outbreak!

If data comes from many small controlled experiments inference is quite easy:

Estimation from many small outbreaks

Example: suppose we have many (n) units of size 2 in which one was initially infected

If m out of the n households resulted in the second individual getting infected then we estimate the transmission probability p by the observed fraction of units in which infection took place:

$$\hat{p} = \frac{m}{n}$$

Note: Parameter estimates are equipped with "hat" (so $\hat{p}$ is an estimate of p)

Estimation from many small outbreaks

If units are isolated (independent) we have a binomial experiment and can easily give confidence bounds:

$$\hat{p} \pm \lambda_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

where $\lambda_{\alpha/2}$ is normal distribution quantile:

95% confidence interval ($\alpha = 0.05$) gives $\lambda_{\alpha/2} = \lambda_{0.025} = 1.96$

Exercise 13: Suppose 27 out of 100 units had the second individual infected. Give a 95% confidence interval for transmission probability p

More about small group outbreaks later

Estimation from one large outbreak

Assume a homogeneously mixing community and no preventive measures

From before: in case of a large outbreak and assuming everyone was initially susceptible, the final fraction infected will be close to the positive solution of

$$1 - \tau = e^{-R_0 \tau}$$

Inference other way around: we observe that a fraction $\tilde{\tau}$ got infected. What is R_0?

Rewrite the equation: $R_0 = -\ln(1-\tau)/\tau$

Our estimate of R_0 is given by the corresponding observed value:

$$\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$$

Exercise 14: Estimate R_0 if 20% were infected during an outbreak

Estimation from one large outbreak

This estimate assumed everyone was initially susceptible!

If in fact a fraction r was initially immune we know from before that τ, the fraction *among the initially susceptible* who got infected approximately equals positive solution of

$$1 - \tau = e^{-R_0(1-r)\tau}$$

This leads to the estimate:

$$\hat{R}_0 = -\ln(1-\tilde{\tau})/(1-r)\tilde{\tau}$$

Note: The over all fraction infected equals $\tilde{\tau}(1-r)$

Exercise 15: Suppose as before that 20% were infected during an outbreak, but that only 50% were initially susceptible and the rest were immune. Compute first $\tilde{\tau}$ and then estimate R_0

Estimation of v_c from one large outbreak

It was shown earlier that: $v_c = 1 - 1/R_0$
By observing an outbreak we can hence also estimate v_c (for the same or similar community but not for any community!):

$$\hat{v}_c = 1 - \frac{1}{\hat{R}_0} = 1 - \frac{\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

If a fraction r was immune in the observed outbreak and $\tilde{\tau}$ of the initially susceptibles were infected this changes to

$$\hat{v}_c = 1 - \frac{1}{\hat{R}_0} = 1 - \frac{(1-r)\tilde{\tau}}{-\ln(1-\tilde{\tau})}$$

Estimation of v_c from one large outbreak

If vaccine not perfect but efficacy E known v_c estimated by

$$\hat{v}_c = \frac{1}{E}\left(1 - \frac{1}{\hat{R}_0}\right) = \frac{1}{E}\left(1 - \frac{(1-r)\tilde{\tau}}{-\ln(1-\tilde{\tau})}\right)$$

Exercise 16. Suppose as previous exercise that 20% of the community got infected but the initial fraction susceptible was 50% (so 40% of these susceptibles were infected). Estimate the critical vaccination coverage for a vaccine having 90% efficacy.

Initial growth rate ρ

For new (so-called *emerging diseases*) and/or lethal diseases it is of course not desirable to wait until the outbreak is over in order to estimate R_0 and other parameters

From before we know $I(t) \approx e^{\rho t}$

So if we observe $I(t_1), \dots, I(t_k)$ it follows that

$$\frac{I(t_k)}{I(t_1)} \approx e^{\rho(t_k - t_1)}$$

It is also true that the cumulative number infected $(n - S(t))$ grows exponentially at the same exponential rate ρ, so it follows that

$$\frac{n - S(t_k)}{n - S(t_1)} \approx e^{\rho(t_k - t_1)}$$

Initial growth rate ρ

This can be used to estimate ρ from data:

$$\ln((n - S(t_k))/(n - S(t_1))) \approx \rho(t_k - t_1)$$

$$\Longrightarrow \hat{\rho} = \frac{\ln((n - S(t_k))/(n - S(t_1)))}{t_k - t_1}$$

(A more proper estimate would be based on logistic regression. Still, this estimator will be biased for various reasons, e.g. time discretization)

Exercise 17: Suppose the incidence ($\approx I(t)$) was observed the first three weeks and the numbers were: 7, 29 and 121 respectively. Estimate ρ.

Estimation of R_0 from initial phase

Suppose we could estimate the growth rate ρ from an emerging outbreak

How about estimating R_0?

Unfortunately the connection between ρ and R_0 is weak (see next slide)

Information about latency period L and infectious period I also needed to estimate R_0

Estimation of L and I hard for two reasons:
1) These periods are rarely observed
2) Even if they were: during the early stages of outbreak short periods are over-represented

Illustration that R_0 and ρ not very related

Illustration. Consider a disease with contact intensity $\beta = 2$ contacts per week and mean infectious $\nu = 1$ week. Then $R_0 = \beta\nu = 2$ and some exponential growth rate ρ.

Consider now another disease having $\beta = 1$ and $\nu = 2$ (less infectious but longer infectious period). Clearly this new disease also has the same $R_0 = \beta\nu = 2$. How about ρ?

The latter is half as fast $\Longrightarrow$ new ρ is half of the former: $\rho_{\text{new}} = \rho_{\text{old}}/2$. So same R_0 but different ρ

However, branching process theory connect ρ and R_0 be means of the **generation time distribution**!

Generation time distribution

Most important property, probably R_0: quantifies how many new infections (on average) infected people cause in the beginning of an epidemic outbreak

Second most important, if we are interested in time evolution (or if we want to estimate R_0 from reported incidence over time!): the **generation time** G

G is the time between getting infected and infecting a new person

$\Longrightarrow$ individuals who infect more than one individual generate several generation times (and individuals who infects noone generate no generation times)

Generation times are not all the same, so G is a **random variable**

$g(s)$ denotes the **generation time distribution** for G. E.g. $g(s)$ is Normal or *gamma distribution* with a given mean and st.d.

Initial growth rate

The growth rate parameter ρ is called the **Malthusian parameter** and depends both on R_0 and the generation time distribution $g(s)$.

Branching process theory: ρ is the solution to the Euler-Lotka equation

$$R_0 \int_0^\infty e^{-\rho s} g(s) ds = 1$$

So if we know the generation time distribution $g(\cdot)$ we can estimate R_0 from observing the exponential growth ρ!

(Problems with estimating $g(s)$ and its consequences a few slides ahead)

Exercise 17.b: Show that if $g(s) \sim \Gamma(\alpha, \beta)$ then Euler-Lotka gives that

$$R_0 = \left(\frac{\rho}{\beta} + 1\right)^\alpha$$

Covid-19: R_0 estimates, **first wave** (original strain)

Covid-19: A common estimate is that $g(s) \sim \Gamma$ with mean 6.5 days and s.d. 4 days. We assume this to apply to all countries!

We estimate "country" specific ρ from reported cumulative case fatalities: starting first day with > 50 cumulative case fatalities (c_0) and two weeks later c_{14} case fatalities: $\hat{\rho} = \ln(C_{14}/C_0))/14$ (Data: Worldometer)

Common dates: first half of March to end of March (before effects of lockdown)

When 50 have died, between 5 000 and 20 000 had been infected so not VERY early in epidemic which is usually atypical and faster (except Norway and Denmark: start instead when > 10 have died)

Covid-19: R_0 estimates, cont'd

Country	C_0	C_{14}	$\hat{\rho}$	$\hat{R}_0$	$\hat{h}_C$
"Norway"	12	89	0.14	2.2	54%
"Denmark"	13	161	0.18	2.6	62%
"Sweden"	62	687	0.17	2.5	60%
"Germany"	68	1275	0.21	3.0	67%
"Belgium"	67	1283	0.21	3.0	67%
"UK"	65	2043	0.25	3.5	71%
"Spain"	55	3647	0.30	4.3	77%

($h_C =$ critical vaccination coverage for herd immunity, more later)

$\Longrightarrow$ There is not one correct R_0 for covid-19!!

Big differences also within countries!
(Sweden starting when > 10 had died gave $\hat{R}_0 = 3.1$)

Problems with estimating $g(s)$ and its consequences

Details: see Britton & Scalia Tomba (2019)

How estimate generation time distribution $g(s)$?

Answer: **Contact tracing**: For some identified cases, it is traced by whom and when they were infected

This gives some observed generation times $g_1, \ldots, g_k$. This is often only way, but problematic:

- Generation time defined forward in time but contact tracing backward in time. Problematic?
- For some cases a unique infector and infection time is identified, but for some there are several possibilities (and some have none)
- onset of symptoms more common to observe than infection times
- Identified cases are often severe cases. Do mild/asymptomatic cases have same generation times?

Toy example

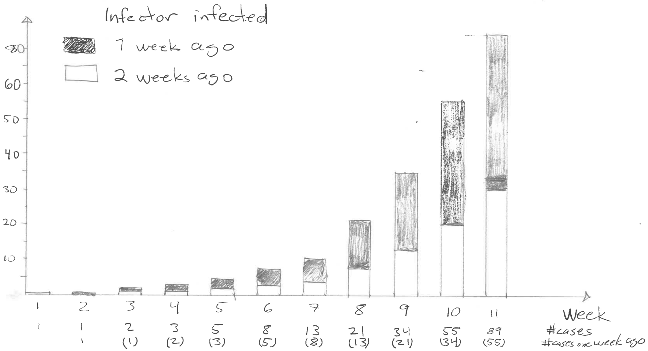
Suppose that $R_0 = 2$, and each infected infects one individual after 1 week and one individual after 2 weeks ($g(1) = g(2) = 0.5$)

What is $E(G)$?

Toy example

Suppose that $R_0 = 2$, and each infected infects one individual after 1 week and one individual after 2 weeks ($g(1) = g(2) = 0.5$)

What is $E(G)$? 1.5 weeks, and $st.d.(G)$? 0.5 weeks (below plot of # infections each week)



Looking backwards: contact tracing

Fibonacci numbers and the Golden ratio ...

$\Longrightarrow$ The mean generation time when contact tracing will be < 1.5

So if you estimate $E(G)$ (or all of G) from contact tracing you will *under-estimate* $E(G)$

Generation times vs Serial intervals

Serial intervals instead of generation times

(We now forgetproblem of looking backwards)

Infection times are hardly ever observed, but onset of symptoms are

$G =$ time between infection times (unobserved)

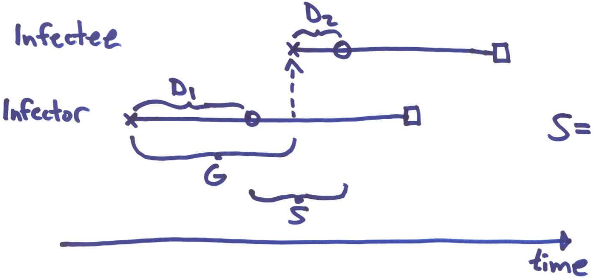
$S =$ time between onset of symptoms (observed)

Generation times vs Serial intervals, cont'd

Generaton times vs Serial intervals

x = infection
o = onset of symptoms
□ = recovery/death

D_1 & D_2: incubation periods
G: generation time
S: serial interval



$S = G + (D_2 - D_1)$

Generation times vs Serial intervals, cont'd

$\Longrightarrow S = G + (D_2 - D_1)$ (D_1 and $D_2 =$ incubation periods of infector and infectee)

So, if incubation times are independent and independent of G, then

$E(S) = E(G)$, and $V(S) \geq V(G)$

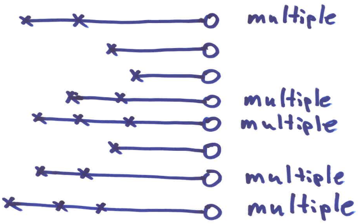
(The relation holds true for all (?) epidemic models)

So, if we estimate $G \sim \{g(s)\}$ from observations on Serial intervals we will *over-predict* variance of G

Multiple exposures

Another problem when contact tracing is that sometimes there are several potential infectors (see illustration on next slide)

Relative infection times of potential infectors



If "multiples" are removed ⇒ Remaing times are shorter

Multiple exposures

If observations with more than one infected are neglected, remaining intervals are biased from below.

This will also lead to *under-estimation* of $E(G)$

Conclusions: looking backwards and neglecting multiple exposures lead to **under-estimation** of $E(G)$ and observing serial intervals rather than generation intervals lead to **over-estimation** of $V(G)$

We now see how this can affect estimates of R_0

Effects of bias in estimates of $g(s)$

$I(t)$ = incidence day t = # infected day t (now discrete time)

How many that get infected day t depends on: R_0 =, basic reproduction number and $\{g(s)\}$ = Generation time

– how many that got infected s days ago? Answer: $= I(t-s)$

Model definition (common model)

$$I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g(s) I(t-s)\right), t = 1, 2 \ldots, \qquad (*)$$

"Pois( )" means Poisson distribution, and the mean equals the parameter, $R_0 \sum_{s=1}^{t} g(s) I(t-s)$

Exercise 17.c: Show that this is more or less identical to the Euler-Lotka equation (Hint: replace the Poisson random variable by its mean)

Effects of bias in estimates of $g(s)$ (cont'd)

$$I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g(s) I(t-s)\right), t = 1, 2 \ldots, \qquad (*)$$

If $\{g(s)\}$ known (or estimated), Eq. $(*)$ can be used for:

1: Estimating R_0 (from observed incidence $I(1), \ldots, I(t)$), or
2: Predicting outbreak incidence $I(1), \ldots, I(t)$ (if R_0 known before-hand)

Both 1 and 2 require knowledge about $\{g(s)\}$

Main question: How to estimate generation time distribution $\{g(s)\}$ and what happens to estimates of R_0 (or predictions $I(1), I(2), \ldots$) if $\{g(s)\}$ is estimated incorrectly?

Effects of bias in estimates of $g(s)$ (cont'd)

Recall, $I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g(s) I(t-s)\right)$
where $I(0), \dots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \dots, I(t)$) affected by the generation time distribution $\{g(s)\}$?

Effects of bias in estimates of $g(s)$ (cont'd)

Recall, $\quad I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g(s) I(t-s)\right)$

where $I(0), \ldots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \ldots, I(t)$) affected by the generation time distribution $\{g(s)\}$?

It is easy to show that the mean parameter

$R_0 \sum_{s=0}^{t} g(s) I(t-s)$ **increases** if:

– $g(s)$ is replaced by $\hat{g}(s)$ which has smaller mean

– $g(s)$ is replaced by $\hat{g}(s)$ which has same mean and larger variance

Effects of bias in estimates of $g(s)$ (cont'd)

Recall, $\quad I(t) \sim \text{Pois}\left(R_0 \sum_{s=1}^{t} g(s) I(t-s)\right)$

where $I(0), \ldots, I(t)$ grows, typically exponentially

How are estimates of R_0 (or predictions $I(1), \ldots, I(t)$) affected by the generation time distribution $\{g(s)\}$?

It is easy to show that the mean parameter

$R_0 \sum_{s=0}^{t} g(s) I(t-s)$ **increases** if:

– $g(s)$ is replaced by $\hat{g}(s)$ which has smaller mean

– $g(s)$ is replaced by $\hat{g}(s)$ which has same mean and larger variance

So, if our estimate of $\{g(s)\}$ has mean biased from below we will **under-estimate** R_0

And if we estimate $\{g(s)\}$ by something with the correct mean but larger variance we will **under-estimate** R_0

Effects of bias in estimates of $g(s)$ (cont'd)

A few slides back we showed three problems when estimating $g(s)$ from **contact tracing**:

1) Looking backwards rather than forward in time: $g(s)$ was biased from below ($E(G)$ under-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

2) What if multiple infector candidates: $g(s)$ was biased from below ($E(G)$ under-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

3) Observing Serial intervals instead of Generation times $g(s)$ has too large standard deviation ($V(G)$ over-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

Effects of bias in estimates of $g(s)$ (cont'd)

A few slides back we showed three problems when estimating $g(s)$ from **contact tracing**:

1) Looking backwards rather than forward in time: $g(s)$ was biased from below ($E(G)$ under-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

2) What if multiple infector candidates: $g(s)$ was biased from below ($E(G)$ under-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

3) Observing Serial intervals instead of Generation times $g(s)$ has too large standard deviation ($V(G)$ over-estimated)
$\Longrightarrow R_0$ will be **under-estimated**

Conclusion: Unless taken account for, all three problems make R_0 *under-estimated*. See Britton & Scalia-Tomba (Interface, 2019)

Biases for Ebola and COVID-19

For Ebola 75% of contacts had multiple potential infectors. The combinded under-estimation of R_0 was $\approx 23\%$

For Corona (Covid19) there was no information of multiple infectors (but I am sure there were!), so only considering bias from backward tracing we believe R_0 is under-estimated by $\approx 12\%$.

Endemic diseases

Consider an *endemic disease* and that $\tilde{s}$ observed

$\tilde{s}$ = average fraction of susceptibles = average relative time spent in susceptible state = average age at infection/average life-length

From before we know $\tilde{s} \approx 1/R_0$

$\Longrightarrow \hat{R}_0 = \frac{1}{\tilde{s}}$

By only knowing the typical infection-age and life-length gives estimate of R_0!

Endemic diseases: estimation of v_c

Same data: $\tilde{s} =$ average age of infection divided by average life-length ($=$ average fraction susceptible in community)

We know that $v_c = 1 - 1/R_0$ (or $v_c = E^{-1}(1 - 1/R_0)$ if vaccine has known efficacy E)

$\Longrightarrow \hat{v}_c = \frac{1}{E}(1 - \tilde{s})$

Exercise 18 Suppose (as with measles) average age of infection is 5 years and average life-length is 75 years. Estimate R_0 and v_c assuming a vaccine having efficacy $E = 0.95$. (How about if $E = 0.90$?)