Stockholms universitet

L3, Inference on stochastic epidemic models

Tom Britton

July, 2019

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?

Stockholms universitet

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

Stockholms universitet

Estimation from one large outbreak

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

Stockholms universitet

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.

Stockholms universitet

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 $I(t_1), \ldots, I(t_k)$ it follows that

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

Initial growth rate ρ

This can be used to estimate ρ from data:

$$\ln(I(t_k)/I(t_1)) \approx \rho(t_k - t_1)$$

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

(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.

Stockholms universitet

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 week (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

Stockholms universitet

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 twice as slow $\Longrightarrow$ new ρ is half of the former:
$\rho_{\text{new}} = \rho_{\text{old}}/2$

Stockholms universitet

Pitfalls when estimating in emerging outbreaks

$I(t)$ = incidence day t = # infected day t

How many that get infected day t depends on:

– R_0 = basic reproduction number = mean number of infections by one infector

– $\{g(s)\}$ = Generation time = typical length between getting infected and infecting new people

– how many that got infected that long back in time = $I(t-s)$

Model definition (common model)

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

$G \sim g(s)$ = generation time distribution: probability that an infector infects a random infectee s days after own infection

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (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?

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

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

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 $\{g(s)\}$?

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

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

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

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

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

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 (or **over-predict** the future outbreak)

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

Pitfalls when estimating in emerging outbreaks (cont'd)

How to estimate generation time distribution $\{g(s)\}$?

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

How to estimate generation time distribution $\{g(s)\}$?
Contact tracing: look up infectors of infected people and compare onset of symptoms

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

How to estimate generation time distribution $\{g(s)\}$?
Contact tracing: look up infectors of infected people and compare onset of symptoms

Three problems with this:

1) Looking backwards rather than forward in time
2) What if multiple infector candidates
3) Serial intervals instead of generation times

Stockholms universitet

Pitfalls when estimating in emerging outbreaks (cont'd)

How to estimate generation time distribution $\{g(s)\}$?
Contact tracing: look up infectors of infected people and compare onset of symptoms

Three problems with this:

1) Looking backwards rather than forward in time

2) What if multiple infector candidates

3) Serial intervals instead of generation times

Conclusion: Unless taken account for, all three problems make R_0 *under-estimated* (or over-predict future outbreak). See Britton & Scalia-Tomba (2018) for details

Stockholms universitet

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!

Stockholms universitet

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 $E = 0.90$?)