L8, More on inference from big outbreaks

Tom Britton

July, 2016

Repetition: Inference from large outbreaks

From lecture 3: basic reproduction number R_0 and critical vaccination coverage v_c were estimated by:

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

if outbreak takes place in a fully susceptible homogeneous community resulting in a fraction $\tilde{\tau}$ getting infected during the outbreak

How about uncertainty?

Uncertainty of previous estimate

Intuition: The larger community (and more getting infected) the less uncertainty

It was mentioned that final number infected $n\tilde{\tau} = Z$ in case of a major outbreak is normally distributed with mean $n\tau^*$ and standard deviation $\sqrt{n\sigma^2}$ where σ^2 depends on model parameters and shown two slides ahead

This result can be used to show that $\hat{R}_0$ and $\hat{v}_c$ are normally distributed with correct means (i.e. R_0 and v_c respectively) and standard errors to be derived using δ-method

The δ-method

Suppose random variable X has mean $\mu = E(X)$ and variance $V(X)$

Then the δ-method gives the following approximation for the mean and variance of $f(X)$, where $f(x)$ is a "nice function":

$$E(f(X)) \approx f(\mu) \qquad V(f(X)) \approx (f'(\mu))^2\, V(X)$$

The approximation holds better the smaller variance X has (i.e. smaller $V(X)$)

The δ-method for $V(\hat{R}_0)$

Probabilists have proven that the asymptotic variance of $\tilde{\tau}$ equals:

$$V(\tilde{\tau}) \approx \frac{1}{n} \frac{\tau(1-\tau)}{(1-(1-\tau)R_0)^2} \left(1 + c_v^2(1-\tau)R_0^2\right)$$

where τ and R_0 are the true parameter values related by $R_0 = -\ln(1-\tau)/\tau$, and c_v is the coefficient of variation of the infectious period.

We now apply the δ-method on $\hat{R}_0 = -\ln(1-\tilde{\tau})/\tilde{\tau}$, we hence have the function $f(x) = -\ln(1-x)/x$

After some algebra we get $V(\hat{R}_0) \approx \frac{1}{n\tau(1-\tau)} \left(1 + c_v^2(1-\tau)R_0^2\right)$

For a standard error estimate we take square roots and replace unknown quantities with there estimates/observed values. The result, also for $\hat{v}_c$, is given by:

Uncertainty of previous estimate

$$s.e.(\hat{R}_0) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\tilde{\tau}(1-\tilde{\tau})}/n}$$

$$s.e.(\hat{v}_c) = \sqrt{\frac{1 + c_v^2(1-\tilde{\tau})\hat{R}_0^2}{\hat{R}_0^4\tilde{\tau}(1-\tilde{\tau})}/n}$$

$c_v^2 = V(I)/(E(I))^2$= squared coefficient of variation of infectious period of individuals (variance divided by the squared mean)

Larger *n* gives smaller standard deviation (as expected)!

Uncertainty of previous estimate

c_v^2 cannot be estimated from final outbreak size – possibly known from before

If not one has to insert a "conservative" bound. E.g. $c_v^2 = 1$: very rarely is standard deviation larger than mean

Exercise 25 Suppose that 239 out of 651 individuals in an isolated village were infected during an outbreak. Estimate R_0 and v_c and give 95% confidence interval for the estimates. Consider both the case when all individuals have the same length of infectious period (so no variation) and the case where its standard deviation is equal to the mean.

Exercise 26 Do the same thing assuming 2390 out of 6510 got infected.

More detailed data

Suppose that disease incidence is observed during outbreak – not only final number

Intuition: more detailed data should improve estimation

Answer: yes, in a couple of ways:

- estimate of R_0 and v_c becomes more complicated, but standard errors are (moderately) smaller
- enables estimation of more parameters: exponential growth rate ρ, latent and infectious period distributions, ...
- possible to detect deviations from model: changing behavior, non-homogeneity, ...

If also information about contacts are available: "transmission probability upon contact" can be estimated

Multitype epidemics

Suppose final size of a multitype epidemic observed: $\tilde{\tau}_1, \ldots, \tilde{\tau}_k$,
$\tilde{\tau}_i =$ observed proportion infected among i-types

Also assumed that community fractions $\pi_1, \ldots, \pi_k$ known.

We want to estimate R_0 which is largest eigenvalue of next generation matrix M

First estimate M. Impossible!! Data has dimension k and M has dimension k^2.

$\Longrightarrow$ M and R_0 cannot be estimated consistently!

Multitype epidemics, cont'd

Why? We can observe who was infected but not who "caused" the infections

Susceptibility easier to estimate than infectivity!

$\Longrightarrow$ only possible to obtain bounds on R_0: lower bound assuming all infections caused by least infected type – upper bound assuming all infections caused by most infected type

Inference in networks

Inference can be performed without an outbreak: estimation of network properties: $E(D)$, $V(D)$, clustering c, ...

R_0, potential outbreak size τ and v_c can then be estimated as a function of transmission probability p

Typical conclusion: Outbreaks are only possible for a disease having higher transmission probability than $p = 0.13$

Or: An STD with $p = 0.08$ can only become endemic in core-groups with average number of partners higher than $E(D) = 4.2$ per year

Inference in more complicated models

More complicated model $\Longrightarrow$ harder inference and more detailed data need

Inference of spread of infections extra hard:

- There are strong dependencies because infections are not independent events (likelihood complicated)
- Many things unobserved: infectious contacts, latent period, infectious period, ...

Inference with more detailed data gives higher precision

Illustration

Suppose an infected infects each susceptible independently with prob p

Data = epidemic chain: $1 \to 2 \to 2 \to 0$

Initially 1 index and 9 susceptible

Likelihood: $L(p) =$
$$\binom{9}{2} p^2 (1-p)^7 \cdot \binom{7}{2} \left(1-(1-p)^2\right)^2 \left((1-p)^2\right)^5 \cdot \binom{5}{0} \left((1-p)^2\right)^5$$
Maximum-likelihood (ML) estimate $\hat{p}$ maximizes $L(\cdot)$:
$\Longrightarrow$ quite easy for a computer

If we instead only know that 5 out of 10 were infected likelihood is much more complicated (a sum over all possible chains)

Alternative approach for complicated models

Basic idea: If likelihood complicated for available data we can "pretend" as if we had more detailed data, estimate parameters under this assumption, recompute some likely more detailed data, re-estimate parameters, ...

This is underlying idea in both EM-algorithm and recently very popular *MCMC*

MCMC: here parameters are treated as outcomes of random variables (Bayesian framework) and even very complicated likelihoods (posterior probabilities) can be evaluated numerically with arbitrary high precision

MCMC: Very computer intensive. Treated specifically in other Modules

Leaky and All-or-Nothing vaccines

Recall from before that

Leaky vaccines: each vaccinee has risk of transmission at each contact (as long as the person is not yet infected) reduced from p to $p\theta$

All-or-nothing vaccines: each vaccinee has probability $1-\theta$ of being fully immune and probability θ of not having any reduced susceptibility at all

Both vaccines have $VE_S = 1-\theta$ and $v_c = \frac{1}{1-\theta}\left(1 - 1/R_0\right)$

Are they equally good vaccines (for same θ)?

Leaky and All-or-Nothing vaccines

It can be shown that P(to get infected within k contacts) is equal for $k = 1$ but for $k > 1$ probability is always larger for leaky vaccine

$\Longrightarrow$ distribution of final size Z satisfies $Z^{(no)} \geq Z^{(leaky)} \geq Z^{(AoN)}$

So, if a community is vaccinated, but not enough for herd immunity, then All-or-nothing vaccine reduces transmission the most (and is hence superior)

If vaccine effect is unknown leaky is worst case scenario

So, an estimate of VE_S is larger if vaccine is assumed to be leaky

Consequence: if VE_S estimated assuming AoN when it really is leaky, the estimate is *under-estimated*