Block 3: Additional methods and dynamics

(stochastic models, health economics, phylodynamics)

Practical notes

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 17: Setting up stochastic models of outbreaks

Practical

Introduction and Objectives:

This practical is divided into three parts. The first part illustrates how you might set up stochastic models in Excel using method 1 presented in the lecture; the second part illustrates how the size of the probability of an effective contact between individuals determines the distribution of outbreak sizes; part 3 illustrates how this model can be extended to describe complex mixing patterns in populations.

By the end of this practical you should:

- be able to set up a simple model in Excel to simulate an outbreak using method 1 (the "Reed-Frost model");
- understand how the size of the probability of an effective contact between individuals influences the outbreak distributions;
- understand how the Reed-Frost model can be extended to describe contact patterns between many different subgroups in a population.

PART I: Setting up a model using method 1

Revision of the Reed-Frost model ("method 1")

As you saw in the lecture, this method simulates the transmission dynamics by tracking the infection process explicitly for each individual in the population and requires you to repeat the following steps in succession:

a) draw random numbers for each susceptible individual
b) compare each random number against the risk of effectively contacting at least one case to determine whether or not he/she became infected.
c) count up the number of resulting cases.

If there are I_t infectious individuals in the population, the risk of effectively contacting at least one case is given by the Reed-Frost equation:

$$\lambda_t = 1-(1-p)^{I_t},$$

where p is the probability that 2 specific individuals come into effective contact per time step.

This equation follows from the facts that

1. The risk of effectively contacting at least one case equals 1 – {the probability that an individual avoids contact with all I_t *cases*}.
2. (1-p) is the probability that an individual avoids contact with 1 case per time step, (1-p)×(1-p) is the probability than individual avoids contact with 2 cases, and so $(1-p)^{I_t}$ is the probability that an individual avoids contact with each of the I_t cases.

In this practical, we will consider a population comprising 10 susceptible individuals (e.g. a hospital ward or a small class) following the introduction of one infectious case. As in the lecture, we will use time steps of 1 serial interval.

Setting up the Reed-Frost model

1. Open up the spreadsheet reedfrost.xls. You should see a sheet with some yellow, orange and blue cells.

The yellow cells contain the key input parameter into the model, namely the probability of an effective contact between 2 specific individuals per time step (i.e. per serial interval). This probability is currently set to be 0.05.

The orange cells (columns B-L) describe the status of each individual in the population at the start, with cells B32, C32, D32 etc reflecting the status of individuals 1, 2, 3 etc respectively at the start.

The first three columns (columns M-O) of the blue cells sum up the number of individuals who are susceptible, infectious or immune in each generation. Column P contains equations for the risk that an individual effectively contacts at least one infectious case, according to the Reed-Frost equation (note that the Excel expression of an equation of the form a^c is just a^c).

Q1.1 What is the risk of becoming infected during the first time step?

2. Change the status of individual 3 to a case by typing the word “case” (without inverted commas!) into cell D32. You should notice that the colour of the cell changes from orange to red.

Q1.2 How does having 2 cases in the population affect the risk of becoming infected? How does the risk change if there are 3 or more cases in the population?

3. Before continuing, change the status of all the individuals (apart from individual 11) to “susceptible” again (by typing in the text “sus” into the corresponding cell).

We now need to draw random numbers to determine what happens next.

4. Click the button with the + sign above column AB.

You should now see some pink cells in columns Q-AA. These cells contain random numbers, which are drawn for a given individual only if that individual is susceptible. For example, in cell Q32, a random number is drawn only if individual 1 is susceptible at the start; otherwise the text “non sus” is inserted in the cell. (See the Appendix of this practical if you wish to revise “if” statements in Excel. Similarly, in cell R32, a random number is drawn only if individual 2 is susceptible at the start; otherwise the text “non sus” is inserted in the cell.

5. Press the F9 key. You should notice that the random numbers in cells Q32-AA32 change. In fact, Excel automatically updates the random numbers whenever anything in the spreadsheet changes.

We now need to use these random numbers to determine the status of all the individuals in the next time step.

6. Click on the + sign to the left of row 44. Then click on cell B33.

Q1.3 What does the command in this cell do? What do you think the status of individual 1 would be at time step 1 if he/she was a case or immune at the start?

7. Check that the status of the individuals is consistent with your answer to the last question.

8. Click on the + sign to the left of row 43.

You should now see all the formulae set up for all the time steps in the model.

We will now investigate how the size of the probability of an effective contact determines the size of the outbreak.

PART II: The effect of p on the outbreak size and number of cases seen in the first generation

Simulating outbreaks

1. Select rows 2 and 26 together, click with the right mouse button and choose the unhide option.

You should now see a figure plotting the number of susceptible, infectious and immune individuals over time. You will also see some blue cells summarizing the outbreak size, and the number of new infections in the first generation.

Q2.1 According to the cell labelled "total in outbreak", what is the current outbreak size?

2. Press F9 to "run" another simulation.

Q2.2 How does your outbreak size change?

3. Change the value for p to take the following values and for each value of p, press the F9 key several times.

i) 0.05 ii) 0.1 iii) 0.25 iv) 0.5

Q2.3 How does the outbreak size change as you increase the size of p?

Q2.4 What happens to the outbreak size if you increase the number of infectious individuals present at the start?

As you saw in blocks one and two, in a deterministic model, the number of new infections per time step is given by the expression:

Risk of infection between time t and t+1 (λ_t)

×

Number of susceptible individuals at the start of the time step (S_t)

4. Use this expression to set up an appropriate equation for the expected number of cases which should be seen in the first generation in cell D25.

5. Change the value for p to be 0.1, making sure that only one case is present in the population at the start.

Q2.5 How many cases in the first generation would we expect to see on average?

Q2.6 How is this number related to the basic reproduction number?

6. Run the model several times for a given value of p.

Q2.7 How does the actual number of cases seen in the first generation in the model compare against that expected?

When working with stochastic models, it is usually helpful to collect the results from many simulations and plot the frequency distributions of the outcome of interest such as the outbreak size. This allows us to determine useful statistics, such as the probability of a large outbreak. We shall create such frequency distributions using our spreadsheet.

Analysing the general distributions of outbreak sizes and the number of first generation cases

1. Click on the button with the + sign above column AQ.

You should now see some grey cells, two figures (currently empty) plotting the frequency distributions of outbreak sizes and the number of cases in the first generation, and a button labelled “Run 30 simulations”.

This button is linked to a macro, which runs the model 30 times, and copies the outbreak size and the number of cases in the first generation for each model run to columns AD and AE. The corresponding frequency distributions are calculated in cells AD-AK.

2. If there is a security warning below the ribbon stating “Macros have been disabled”, click on the “Options” button next to this warning and then select the “Enable this content” option, before clicking on OK. Now click on the Click on the button labelled “Run 30 simulations”.

3. Steadily increase the value for p from about 0.05 to 0.5 and each time, run the model 30 times. (You may prefer to hide the columns with the random numbers by clicking on the button with the ‘–‘ sign above column AB, so that you can see the value for p and the frequency distributions simultaneously.)

Q2.8 What kind of outbreaks are most likely if p is i) <0.1 ii) >0.1? iii) equal to 0.1?

Q2.9 Is this what you would expect? Why? (Hint: What is the R_0 when p=0.1?)

4. Look at the distribution of outbreak sizes.

Q2.10 Could you use this to infer what the R_0 might be for a new pathogen such as influenza caused by the H5N1 virus?

Q2.11 Suppose you have distributions of outbreak sizes of measles in communities. What further information would you need to infer the R_0?

PART III: Extending the model to describe contact between subpopulations

The model described above can be extended to describe contact between different population groups.

For example, suppose that there are now two groups (e.g. two wards in a hospital). We might now assume that there are two contact parameters, namely:

1. p_{in} which describes the probability of an effective contact between two specific individuals in the same subgroup
2. p_{out}, which describes the probability of an effective contact between a specific individual in one subgroup with a specific individual in the other subgroup.

Assuming that the two groups are otherwise independent, we can extend the Reed-Frost equation on page 1 to obtain an expression for the probability that a susceptible individual in a given subgroup will be infected:

$$\lambda_t = 1-(1-p_{in})^{I_{in,t}}(1-p_{out})^{I_{out,t}}$$

Here, $I_{in,t}$ is the number of infectious individuals at time t who are in the same group as the susceptible person being considered and $I_{out,t}$ is the number of infectious individuals at time t who are not in same group as the susceptible person being considered.

Q3.1 What do the components $(1-p_{in})^{I_{in,t}}$ and $(1-p_{out})^{I_{out,t}}$ represent?

1. Turn to the worksheet Reed-Frost2.

This worksheet is identical to the model that we were using in Parts 1 and 2 of this practical, except that it comprises two linked subpopulations, with each subpopulation consisting of 11 individuals. p_{out} is assumed to differ by a factor (equal to the value of rel_contact_freq) from p_{in}. This kind of model is technically known as a "stochastic metapopulation" or "patch" model, since each of the subpopulations ("patches") are linked.

2. Run the model for different values of p_{in} and rel_contact_freq.

Q3.2 How does the likelihood of an epidemic in the other population depend on p_{in}?

This kind of model can be easily extended to deal with many different subpopulations.

Q3.3 Suppose we wanted to describe children contacting others in a household, at school and in the wider community. What would be the expression for the risk of a susceptible child becoming infected in each time step?

Further exercises

If you have finished the practical early or if you wish to consolidate your understanding of the concepts covered in this session, please try the following exercises:

1. Supplementary exercises (see the supplementary questions folder on Moodle or in the network folder with the model files) which illustrate how method 2 described in the lecture may be used to set up models.
2. Exercises accompanying models 6.2. 6.3 and 6.4 of the recommended course text[1] (see www.anintroductiontoinfectiousdiseasemodelling.com).

References

1. Vynnycky E, White RG. An Introduction to Infectious Disease Modelling". Oxford University Press, Oxford 2010

Appendix

Revision of "IF" statements in Excel

In Excel, the following command compares the contents of cell X999 against the value of Y, and if the value in cell X999 is less than the value of Y, the letters 'AA' are inserted in the cell you're working in; otherwise the letters 'BB' are inserted:

=if(X999<Y,"AA","BB")

With the following command:

=if(X999<Y,rand(),"BB")

if the value in cell X999 is less than the value of Y, a random number is inserted in the cell you're working in; otherwise the letters 'BB' are inserted

Similarly, the following expression:

if (X999="AA","BB","CC")

compares the contents of cell X999 against the letters 'AA' and if they are identical, the letters 'BB' are inserted in the cell you're working in; otherwise, the letters 'CC' are inserted.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 18: Cost-effectiveness of seasonal influenza vaccination

Practical

Learning objectives

1. To estimate the incremental cost-effectiveness ratio of an infectious disease intervention like vaccination.
2. To conduct both one-way and probabilistic sensitivity analyses.
3. To construct and compare both static and dynamic models of infectious disease interventions.

Overview

The purpose of this practical is to show you what types of data are needed to estimate the incremental cost-effectiveness ratio (or ICER) of introducing a vaccination programme, and to carry out probabilistic sensitivity analysis (PSA) on the results. We will look at the introduction of seasonal influenza vaccination just before the start of the influenza season. This has the advantage that the effects of vaccination take place within a single year, so we don't have to worry about the effects of discounting (apart from one exception, which we will see later).

Modelling seasonal influenza

1. Open up the workbook "CEA.xlsx".

It has six worksheets. On the worksheet "No vaccine", you will see an SIR model of influenza in a population split into two age categories (children and adults). Parameters and outputs affecting children and adults are coloured differently. Some members of both groups are already immune at the start of the influenza season, probably because they were infected by influenza in a previous season. Try to understand the way the model works – you should be familiar with such models by now.

Q1. The worksheet calculates two clinical outcomes: the number of clinical cases and deaths due to influenza at every time step. How are these outcomes calculated?

2. Look at the worksheets "Vaccine strategy A" and "Vaccine strategy B"

Q2. What is the difference between "Vaccine strategy A" and "Vaccine strategy B"?

Q3. Which vaccine strategy has the greatest impact on influenza infections? Why?

Estimating the cost-effectiveness of vaccination

Look at the "Economics" worksheet. You will now fill in this worksheet in order to estimate the ICER of both vaccination strategies (compared to no vaccination and to each other).

The "Children" and "Adults" sections simply collect some key outcomes that will be needed for subsequent health economic calculations. These can be found by referring to appropriate cells in the previous worksheets. For instance, the number of child clinical cases in the "No vaccination" strategy simply refers to cell G24 in the "No vaccine" worksheet.

In the next section labeled "Totals", cells representing the total number of clinical events, costs and QALYs lost associated with influenza (or influenza vaccination) in both age groups. This uses the results in the "Children" and "Adult" tables, and the parameters in the "Parameter" table at the top of the worksheet: these describe the cost of influenza vaccination, cost of treating a clinical case of influenza and QALY loss associated with both clinical cases and deaths due to influenza.

Q4. The QALY loss associated with an influenza death is set to 30. How do you think such a number may be estimated? Can you think of any reason why having a single number for this value may not always be a good idea?

Q5. Now fill in the table under "Differences", as well as the incremental cost per life saved, death prevention and QALY gained below that. This table measures the difference in clinical events, costs and QALYs lost between any two strategies (A vs. no vaccine, B vs. no vaccine and A vs. B). It can be completed simply by referring to appropriate cells in the "Totals" table.

Note: You may want to think about the order in which you subtract one quantity from another. For instance "Clinical cases prevented" under "No vaccine to A" refers to the **decrease** in clinical cases when you introduce vaccine strategy A, compared to when you have no influenza vaccine strategy, i.e. the number of clinical cases without a vaccine minus the number of clinical cases when you have vaccine strategy A. However, "Vaccine costs" under "No vaccine to A" refers to the **increase** in vaccine costs when you introduce vaccine strategy A, compared to when you have no influenza vaccine strategy.

Q6. Suppose this model is being used to inform vaccination decisions in a country with a threshold of £5,000 per QALY gained for a health technology to be cost-effective. Based on the results of your calculation, what would you advise decision makers to do?

Q7. Do you think that this is a realistic model of the economic consequences of influenza vaccination? What could be done to make it more realistic?

Sensitivity analysis

In the previous cost-effectiveness calculation, we assumed that the economic parameters (costs and QALY losses) were known precisely. In this worksheet, we will estimate the impact of uncertainty on the cost-effectiveness of strategy A (compared to no vaccination) using both one-way and probabilistic sensitivity analysis.

Q8. Change the value of the following parameters by ±25%. (Note that the parameters in the "Vaccination strategy A" and "Vaccination strategy B" worksheets are linked to those in the "No vaccination" worksheet, so you only need to alter the parameters in the "No vaccination worksheet" to have the other two worksheets automatically updated.) What happens to the ICER in each case?

Parameter	**Worksheet**	**Cell**	**Lower value**	**ICER**	**Upper value**	**ICER**

Transmission coefficient child-child	No vaccination	D6	0.4125		0.688	
Case-fatality risk in children	No vaccination	C19	0.0075%		0.0125%	
Cost per clinical case	Economics	B3	9		15	

Q9. Which parameter that you varied has the largest impact on the ICER? In what way does (i) decreasing it by 25% and (ii) increasing it by 25% alter the ICER? Why?

Q10. Now let's do a PSA instead of just changing parameters one at a time. Look at the worksheet "PSA". The yellow cells show the maximum and minimum value assumed for the distribution of the four economic parameters in a PSA. Assuming that they are uniformly distributed, the first line under "Scenario" calculates one possible value they can take by sampling from these distributions. How are the values calculated?

Q11. Do you think a different type of distribution may be better than a uniform distribution? How could it be implemented in Excel?

Q12. Copy the first line under "Scenario" down until you produce 100 scenarios. What happens?

Q13. For each scenario, fill in the rest of the table which calculates the costs and QALYs lost associated with no vaccination and vaccination option A, as well as the ICER of vaccination option A. Use this to calculate the mean and 95% interval of the ICER. [Hint: look up the Excel functions AVERAGE() and PERCENTILE()]

Q14. How certain are you that the advice you gave in Q8 above is correct?

Congratulations - you have successfully carried out a cost-effectiveness analysis of influenza vaccination based on a transmission dynamic model!

Optional: Comparing the results to those obtained using a static model

Many cost-effectiveness analyses of vaccination and other infectious disease interventions are based on static models that do not take into account transmission. As a result, these often underestimate the impact that vaccination will have, although in some cases they can also underestimate detrimental effects (like changes to the age profile of a disease).

To see the difference between static and dynamic models, we will use the worksheet "Static model" to construct a static model which is equivalent to the dynamic model except that it does not take into account transmission of influenza. Look at the table in the worksheet, which takes you through the steps to estimate the cost-effectiveness of vaccination strategy A using a static model. The first two lines are filled in for you.

Q15. How are the numbers of clinical infections calculated?

Q16. Use an equivalent formula to calculate the number of deaths you would expect (i) without vaccination and (ii) with vaccination strategy A, using the same static model, and use that to fill in the next row of the table. What assumption did you have to make to do the calculation?

Q17. Now that you know the number of clinical cases and deaths under both vaccine strategies, use the same method as in the previous section to work out the ICER of strategy A (paediatric vaccination) compared to no vaccination.

Q18. Compare the ICERs you obtain using a static and a dynamic model. What do you notice? What does the difference tell you?

Epilogue

Note that the model is greatly simplified and the parameters are not necessarily accurate, so the results of this exercise should not necessarily be used as an indication of how influenza vaccination might work in the real world. If you are interested in more realistic models you may want to take a look at the following papers:

Baguelin M, Camacho A, Flasche S, Edmunds WJ. Extending the elderly- and risk-group programme of vaccination against seasonal influenza in England and Wales: a cost-effectiveness study. BMC Medicine 2015;13:236.

Newall AT, Dehollain JP, Creighton P, Beutels P, Wood JG. Understanding the cost-effectiveness of influenza vaccination in children: methodological choices and seasonal variability. Pharmacoeconomics. 2013;31(8):693-702.

Pitman RJ, Nagy LD, Sculpher MJ. Cost-effectiveness of childhood influenza vaccination in England and Wales: Results from a dynamic transmission model. Vaccine 2013;31(6):927-42.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 20: Setting up discrete time stochastic models in Berkeley Madonna

Practical

Introduction and Objectives

You used Excel in the recent stochastic modelling practical. This is useful for understanding how stochastic models work, and some of the implications of accounting for the role of chance in models, but is not so useful in more realistic modelling situations as Excel is rather limited. In this session we will use Berkeley Madonna to set up a discrete time stochastic (population-based) model, and compare the results to an equivalent deterministic model. Apart from the model that was used in the session on rubella vaccination in block 2, all of the Berkeley Madonna models that you have used so far have used *differential* equations, and not *difference* equations. This practical will therefore provide further experience of using models that are set up using difference equations in Berkeley Madonna.

By the end of this session, you should:

- be familiar with the basic methods for setting up a discrete time stochastic model in Berkeley Madonna.
- gain an understanding of how chance can influence the epidemiology of outbreaks in small populations
- be able to set up a deterministic model using difference equations in Berkeley Madonna.

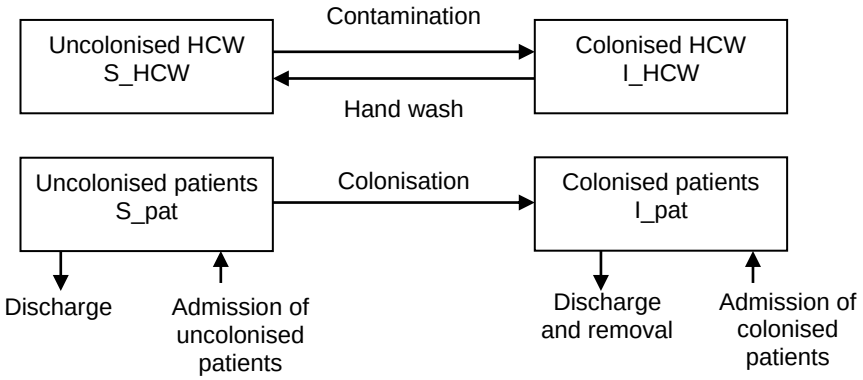
Epidemiological background and summary of model

In this session we will be modelling the spread of methicillin-resistant *Staphlococcus aureus* (MRSA) infection in the hospital environment. As you may be aware, this is a major problem in the UK and many other countries, resulting in large numbers of cases and deaths. The extra complications associated with MRSA infections leads to a greater length of stay in hospital, as well as increased interventions, with the net result that MRSA and other health-care associated infections result in significant costs to the health system.

You will be using a model of the spread of a hand-borne pathogen first published by Ben Cooper and colleagues in 1999. Although the model is general, it was parameterised with MRSA in mind. Details of the model, and how the parameters were estimated are given in the accompanying paper (Cooper BS et al. J Hosp Infect (1999) **43**:131-47). A brief description of the model is given below.

The model considers a single, typical, ward. Within this, two populations are considered: patients and health care workers (HCW). Patients can be either colonised (i.e. carry the organism and be infectious) or uncolonised (i.e. susceptible). In the model used here, they are called S_pat, I_pat (for susceptible and infectious patients respectively). Once patients are colonised, they are assumed to remain in that state (the rate at which they lose the infection is much lower than the rate at which they would be discharged from hospital).

Health care workers can carry the organism on their hands, and are termed here S_HCW and I_HCW, depending on whether they carrying the organism or not. [Note that Cooper et al. use x and y, and x' and y' in their paper for susceptible and infectious patients and HCW respectively. We have modified the symbols a little in an attempt to render the model easier to read]. Patients are admitted and discharged. A proportion of admitted patients are already colonised. Patients are removed via normal discharge, or because they are detected as being colonised and are then effectively isolated (that is they are no longer a source of infection). The wards are assumed to operate at maximum capacity at all time, so any patients that are removed or discharged are immediately replaced. Health care workers are assumed to remain on the ward. In the base case a ward is assumed to consist of 20 patients and 3 HCWs. It is assumed that the patients do not have any epidemiologically relevant interaction (i.e. they do not contact each other). Health care workers contact patients, if the patient is colonised then the HCW can pick up infection on his/her hands and act as vectors for infection to spread. The HCW contact patients randomly on the ward. When HCW wash their hands, they remove the infection and return to the uncolonised (S_HCW) state. The flow diagram of the model is given below (adapted from Figure 1 in Cooper et al. 1999).



The baseline parameter values are given in Table 1 of Cooper et al, and are assumed to represent a typical general medical ward. They are reproduced below.

Parameter	**Value**	**Explanation**
n_pat	20	number of patients
n_HCW	3	number of HCW
mu	0.1	patient removal rate, per day
mu_HCW	14	handwashing rate, per day
gamma	0.1	detection rate of colonised patients
sigma	0.01	proportion of admissions already colonised
c	5	patient/HCW contact rate, per day
p	0.1	HCW-patient transmission probability
p_prime	0.1	Patient-HCW transmission probability
Beta	=c*p	HCW-patient transmission rate
Beta_prime	= c*p_prime	Patient-HCW transmission rate

Setting up the deterministic model

We will start by using a deterministic version of the model. Open up the model Nosocomial_det.mmd. The model set-up should look reasonably familiar to you. The major difference is that the model is implemented as a system of difference equations. That is, the equations now take the following form (using susceptible HCW as an example)

S_HCW at time t+1 = Susceptible HCW at time t
– Transmission
+ Removal of contamination

I.e. The number of susceptible health care workers at time t+1 is equal to the number of susceptible HCW at time t, minus those that have been infected over that time step, plus previously infected HCW who have decontaminated their hands over the time period.

In Berkeley Madonna this is coded as:

```
next S_HCW = S_HCW - TransToHCW+Rem_cont
```

where the transitions (the # people moving from one state to the next) are:

```
TransToHCW = (Beta_prime*I_pat*S_HCW / n_HCW)*DT
```

```
Rem_cont = (mu_HCW*I_HCW)*DT
```

Note that the transitions have to be in the units of the time step (DT), which means multiplying them by DT, as all the parameters were defined as daily rates.

The rest of the model is coded in a similar manner.

Run the model.

Q1 What is the expected number of infected patients and infected HCW after 1 year using the base-case parameter set?

Cooper et al. show three typical time-courses for three different epidemic. These are given in Figure 2 of Cooper et al, and are reproduced below:

Q2 Do you get similar epidemic curves? And if not, why not?

Q3 Do you always get the same epidemic pattern (if you start with the same parameter values and initial conditions)? Does this differ from Cooper et al.'s results and why?

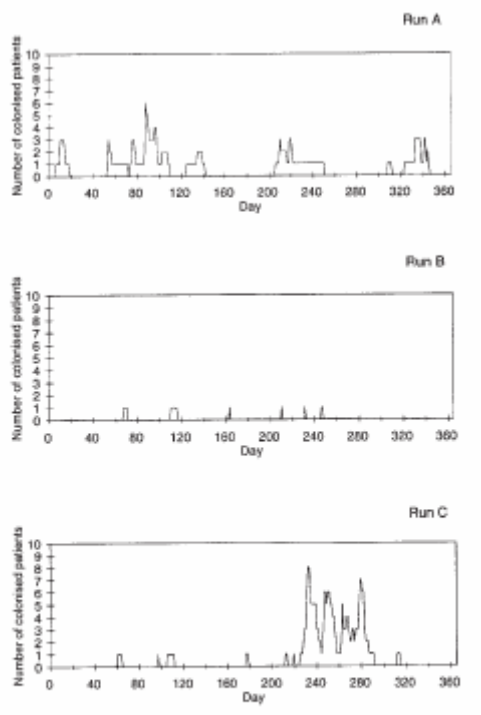


Figure 2 Sample simulation runs from the model, all using the default parameter values (given in table 1) which give an R_0 value of 0·57. These runs were selected to illustrate the high degree of variability that occurs in the stochastic model. In the following graphs, all parameters are set to their default values, unless otherwise stated.

Setting up the stochastic version of the model

The *expected number* of transitions per time step is given by the deterministic version of the model. That is, the expected number of, for instance, infected HCW who decontaminate their hands at a given point in time is (as before) given by:

```
Rem_cont = (mu_HCW*I_HCW)*DT
```

What we need to know is the *actual number* who decontaminate their hands during this time step, given that this process is also affected by chance (that is the actual number is a random variable). We can view each of the infected HCW as a Bernoulli trial, and they either do or do not decontaminate their hands in the time step in question. That is, the number of HCW who do decontaminate their hands would be given by a Binomial distribution with the probability of success given by (mu_HCW*DT), and the number of trials given by the number of infected health care workers (I_HCW). That is, we can replace the deterministic version of the above with:

```
Rem_cont = binomial(mu_HCW*DT,I_HCW)
```

Where the first term in the brackets is the expected probability of success, and the second term (after the comma) is the number of trials.

Similarly, the expected number of transmissions to HCW is given by the deterministic model, i.e.:

```
TransToHCW = (Beta_prime*I_pat*S_HCW / n_HCW)*DT
```

The actual number who may be infected will be given by a binomial distribution with the probability of success given by the deterministic model (i.e. Beta_prime*I_pat / n_HCW * DT) and the number of "trials" given by the number of susceptible HCW (i.e. S_HCW). Again, in Berkeley Madonna code, this would be given by:

```
TransToHCW = binomial((Beta_prime*I_pat / n_HCW)*DT,S_HCW)
```

Q4 Amend the Berkeley Madonna code, converting your discrete time deterministic model to a discrete time stochastic model.

Q5 Compare your results to those that you obtained before (or those of Cooper et al.). Do you now get a similar pattern of epidemics as was recorded in Cooper et al's figure 2 (reproduced above)?

Q6 How does the pattern of outbreaks vary as you increase p?

Supplementary Questions: probabilistic sensitivity analysis

Probabilistic sensitivity analysis (also called *uncertainty analysis*, and sometimes rather loosely referred to as Monte Carlo simulation), is an increasingly common method of taking into account uncertainty in many different parameters simultaneously (it will be covered in a forthcoming lecture). The method works as follows. Instead of defining a single parameter value, parameters of interest (in principle this may include all of the parameters) are given a distribution (sometimes called a prior distribution, or an input distribution). Then the model is run many times, each time a (potentially) different value for the parameters of interest are drawn from these distributions. A distribution of results of the model (output distribution) is then built up.

So, for instance, you might have information that the parameter p is likely to be around 0.15. Usually you would set p=0.1 in your model and then, perhaps, do some sensitivity analysis of the results to this, i.e. you may vary p and check the impact on the results. Instead, however, if you are uncertain about p, you could capture this uncertainty by defining it as a distribution. Let's say that your mean estimate for p is 0.1, with a variance of 0.08. If p is a continuous variable a natural distributional assumption might be that it is a normally distributed random variable. That is, you define p to be normally distributed with a mean of 0.1 and a variance of 0.08. You could follow a similar procedure for other parameters. [*Note that the choice of distribution is critical, and needs to be done carefully. Unfortunately we do not have the time to go into the intricacies of this here.*] You then run the model many times. By keeping the results of all of these runs you are able to build up an output distribution.

There are a number of different ways of doing this in Berkeley Madonna. The one that is described below makes use of the inbuilt statistical functions in Berkeley Madonna (normal, poisson, binomial, uniform, gamma). However there is one trick that you need to employ.

If you simply set a parameter to be drawn from a distribution (e.g. type p = normal(0.1,0.08)) then at every time step a new value for this parameter will be drawn. This is not what you want. You want to draw a value for this parameter once at the beginning of the simulation, and keep it at this value throughout that given model run. You then want to run the model many times – each time drawing a new value for the parameter p (and the other parameters of interest). One way of doing this in BM is to declare your parameter as a variable, initialise your parameter by drawing it from a distribution, and then for the rest of the run, set the parameter to be its initial value.

In BM, therefore, you could type:

```
Init p = normal(0.1,0.08)
Next p = p
```

You could do the same for other parameters.

Then you need to run the model many times and capture the results. One way to do this would be to use a batch-run, indexing your batch-run on a counter that doesn't do anything in the code.

Q7 Return to the original deterministic version of the model and modify it, so that the p is drawn from a normal distribution with mean = 0.1 and variance = 0.08, using the method described above (remember to either comment out, or delete the line of code where p is set to equal 0.1). Press "run" a few times, and check the behaviour of the model.

Q8 Is this model "stochastic" or "deterministic"?

Q9 Perform a batch run 100 times, indexing on the counter "Batchindexnumber" and keep a track of the value for p for each run. That is output p as a graph or table. To do this you may want to output all of the runs to a table (i.e. check the "Keep Runs Separate" box in the Batchruns option). Are all the values of p biologically plausible?

Q10 You may want to modify your distributional assumption for p. You could, for instance, make use of the max function in Berkeley Madonna (Syntax max(x1,x2,..) to prevent p taking a negative value. Try this, and re-run the model a 100 times again, keeping a track of the value of p.

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 21: The applications of Phylodynamics

Practical

Objectives

By the end of this practical you should:

* Understand the relationship between genetic relatedness and transmission events.
* Know how to reconstruct and interpret a phylogenetic tree
* Understand how time and spatial diffusion of an epidemic is inferred from a phylogeny

Introduction

In the lecture we saw how epidemiological processes leave a measurable imprint on pathogen genomes sampled from infected individuals. These processes can be recovered from genetic data sampled at different times and places, using statistical inference methods that take into account the sequences' shared ancestry.

In this practical session, we will reconstruct and interpret the Influenza A H1N1 2009 (H1N1/09) epidemic in England, based on a set of viral sequences isolated during the outbreak.

We will identify transmission chains of H1N1/09, reconstruct their migration patterns, estimate the number of independent introductions in England, and infer the time and geographical location of these introduction events. In order to do so, we will apply the following procedure:

1 – We will reconstruct the phylogeny of influenza genomes sampled in England and other countries where H1N1 infections were diagnosed.

2 – We will apply a molecular clock model to the data in order to fit the phylogeny to real time scales.

3 – We will infer the migration patterns of H1N1 into and within England using Bayesian Markov chain Monte Carlo (MCMC) phylogeographic inference.

All files for this practical can be found either in your H: drive or in the folder "U:\Download\Teach\scmodels\".

The programs used for this practical are in the 'Teaching' folder from within the application window of your workstation.

Section 1 - Phylogenetic reconstruction

We will use the package **MEGA** to reconstruct the phylogenetic tree. MEGA is a multiplatform, graphical interface for phylogenetic reconstructions and analyses of sequence data. MEGA is freely available at *www.megasoftware.net/*. It runs on Windows, Linux and Mac OS.

1 - Sequence data

The file *'H1N1.flu.2009.fas'* contains 50 full-length influenza A H1N1/09 genomes. These nucleotide sequences correspond to influenza viruses sampled in Canada (n = 5), China (n = 3), England (n = 23), Mexico (n = 4), Peru (n = 1), and the United States of America (USA; n = 14) between April and June 2009. The sequences are 12,734 nucleotides long.

1.1. Start MEGA by double-clicking on the corresponding icon (the big M).

1.2. Import the sequence file: *File > Open a File/ Session… > Open a File > H1N1.flu.2009.fas.* You can also use the '**Data**' menu in the top tool bar. If the sequence file doesn't appear in the MEGA dialog box, select 'All files (*.*)' in the '**File of Type**' tab. Select the file 'H1N1.flu.2009.fas'.

1.3. Since the sequences in the file are already aligned, select **'Analyze'** in the dialog box. Select '**Nucleotide Sequences**' in the **'Input Data'** box. Confirm that the sequences are **protein-coding**. Select the **standard** (i.e. universal) genetic code. A new icon ('**TA**') will appear in the window.

Reminder. To infer the phylogenetic relationship of a set of sequences, these need to be 'aligned'. That is, nucleotide positions need to be arranged in columns to ensure that we compare ***homologous*** *positions one to another. In evolutionary biology, homology means similarity due to decent from a common ancestor. Two genes are homologous if they descent from an ancestral gene. Likewise, two nucleotides in different sequences are homologous if they correspond to the same nucleotide position in the ancestral gene. Note that two objects can be 'similar' without being 'identical'.*

1.4. Click on the '**TA**' icon to display the sequences.

The '**Sequence Data Explorer**' window will appear. By default, MEGA only shows nucleotide substitutions, i.e. nucleotides that differ from the most frequent residue at a given position. Identical nucleotides are replaced by a dot ("."). To see the entire alignment, go to '**Display**' in the menu and unselect the '**Use Identical Symbol**' option. Add colours to the cells for a better visibility of the alignment: *Display > Color cells.*

The sequence alignment is now displayed as a matrix, where lines correspond to viral samples and columns to nucleotide positions (see **Figure 1)**. Cells are coloured by type of nucleotides (A, C, G or T) and missing information, or gaps, are indicated by a dash ("-").

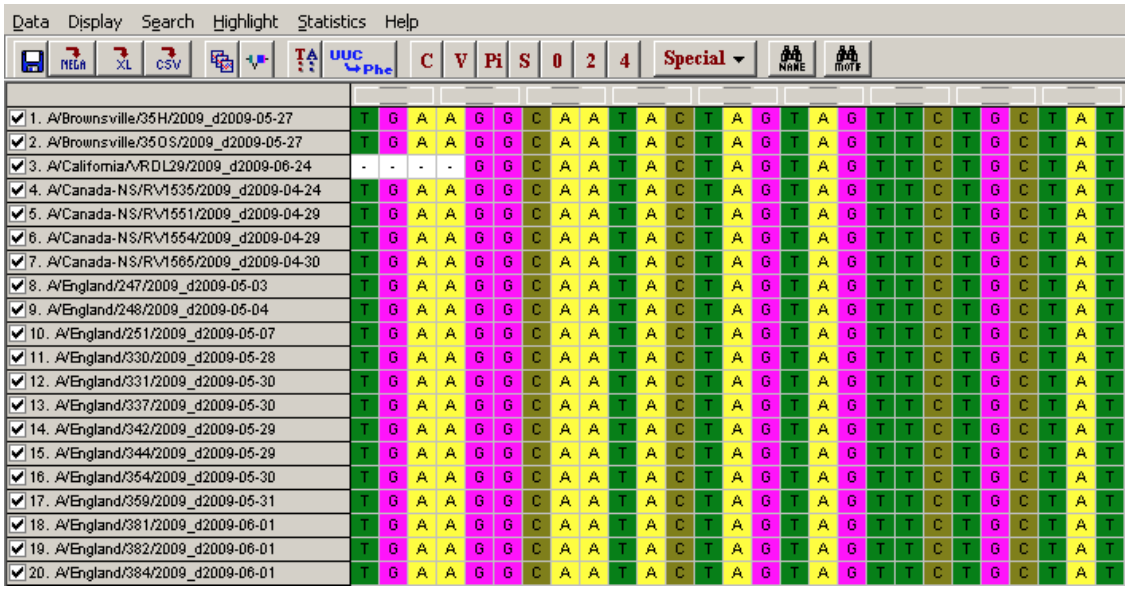


Figure 1. Nucleotide sequence alignment of influenza A H1N1 genomes (detail)

The unique identifier of the sequences is shown at the left hand side of the nucleotide matrix and contains the strain name (e.g. *A/England/247/2009*) followed by the date of sampling (e.g. *d2009-05-31*) in the format dYYYY-MM-DD.

In the matrix, we can to spot 'rare' nucleotide substitutions present in one or more sequence (scroll along the sequence alignment to spot some). These single nucleotide polymorphisms (SNPs) allow us to identify viruses that are genetically related (they share a common SNP) and infer epidemiological linkage between them. If viruses sampled from different individuals have the same SNPs, we can assume that they form a specific strain infecting these individuals. This is the property we use to reconstruct transmission chains from a phylogenetic tree.

2 – Tree reconstruction

MEGA implements several methods for reconstructing trees. For the purpose of this practical we will use a fast, genetic distance-based method called **neighbor-joining** [1].

2.1. Close the **Sequence Data Explorer** and come back to the main window.

2.2. In the **Phylogeny** tab, select '**Construct/Test Neighbor-Joining Tree…**'. An '**Analysis Preference**' window will appear (see **Figure 2**).

This is where we set up the parameters of the phylogenetic reconstruction. Some of these options would requires more time than we can afford to be properly explained, so we will concentrate on the user-defined parameters (highlighted in yellow):

- **Test of Phylogeny**: This option determines the method used to test the reliability of the reconstructed phylogeny. Select '**None'**.
- **Substitution Type**: This parameter determines the genetic unit we will use to infer linkage between sequences (e.g. nucleotides, codon or amino acids). Select **Nucleotide**.

- **Model/Method**: Here you select a stochastic model for estimating evolutionary distances (see lecture). Select '**No. of differences**'. This is the raw number of nucleotide substitutions (or SNPs) two compared sequences differ from.
- **Substitutions to Include**: We will include all types of nucleotide substitution to infer the tree. Select '**d: Transitions + Transversions'**.
- **Gaps/Missing Data Treatment**: This option determines how missing information (i.e. gaps in the alignment) is handled. We will remove all sites containing gaps before the calculation. Select '**Complete deletion'**.
- **Select Codon Positions**: We will use all available information for the tree reconstruction. Make sure **1st**, **2nd**, **3rd** and **Noncoding Sites** are selected.

2.3. Click on '**Compute'**. A tree will appear. It may take up to a few minutes, so be patient.

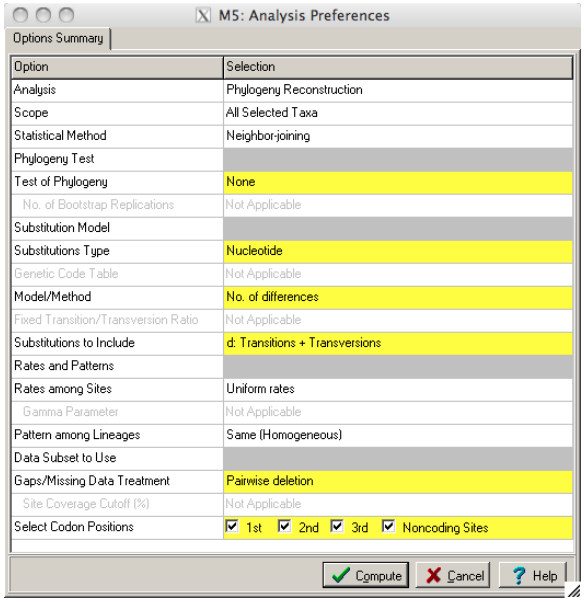


Figure 2. Neighbor-joining analysis preferences

2.4. Save the tree in your folder from the Tree Explorer Window: *File>Export Current Tree (Newick)*

(Name the file as you please, adding the extension '.**tre**' at the end)

3 – Tree interpretation

Take a few minutes to familiarise yourself with the tree.

*Reminder. The components of a phylogenetic tree are shown in **Figure 3**. The sequences, or taxa (plural for taxon), are positioned at the end of the external branches. Related sequences are linked by a node (their most recent common ancestor). Two or more sequences descending from a node form a 'clade' (or cluster). The length of a branch represents the genetic distance between two nodes or between a node and a taxon, i.e. the number of mutations accumulated since divergence. The root corresponds to the common ancestor of all the taxa.*

Since the sequences represent viruses sampled from different individuals, a clade in the tree corresponds to a transmission chain. The number of sequences in a clade reflects the number of infections sampled from that transmission chain. The root of the tree corresponds to the origin of the epidemic.

Q1: How many times was H1N1/2009 introduced in England during the outbreak?

Q2: What is/are the most likely geographical origin of the English strain(s)? (Take the sampling dates into account in your reasoning)

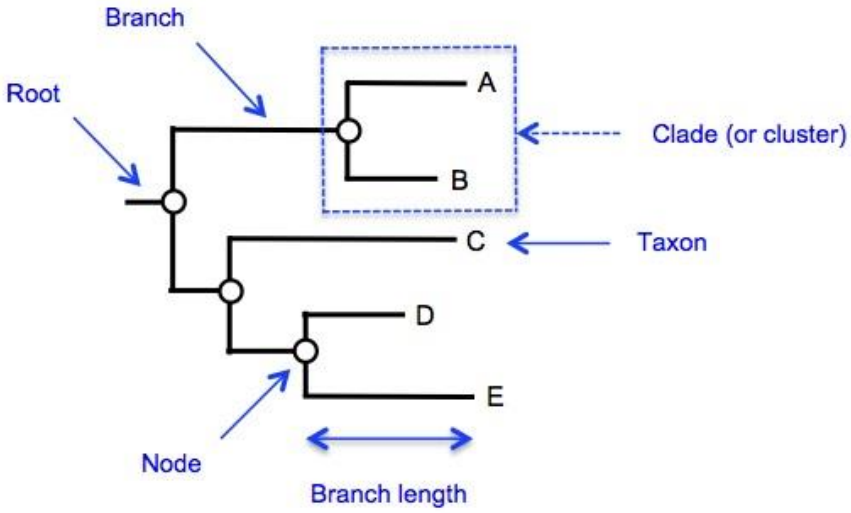


Figure 3. Components of a phylogenetic tree

Section 2 – Dating the introduction(s) of H1N1/09 in England

We will now estimate the time frame of these migration events. This time frame can be inferred from the rate at which mutations are accumulated in gene sequences. If, for instance, two sequences differ by 4 mutations and their rate of evolution is known to be around one substitution per nucleotide position per year, then 2 years have elapsed since they diverged from their common ancestor (i.e. 1 substitution x 2 year x 2 sequences = 4 mutations).

The first step of this procedure involves testing whether the genomes in the tree have evolved at a constant rate over time or not. This is called testing the **molecular clock hypothesis**.

If the molecular clock is constant, or '**strict'**, the genetic distance between two sequences will be proportional to the time since these sequences last shared a common ancestor (as in the example given above). If the molecular clock is not constant, the correlation between genetic distance and time since divergence is weaker. The molecular clock is then said to be '**relaxed'**. Assuming a strict or relaxed molecular clock will have an impact on the dating of phylogenetic nodes. We will therefore test the molecular clock hypothesis before dating the tree.

1- Molecular clock testing

1.1. Put Mega aside and open the software **TempEst** in the 'Teaching' folder.

TempEst [2] is a tool for investigating the 'temporal signal' of molecular phylogenies, i.e. testing whether there is sufficient genetic change between sampling times to reconstruct a statistical relationship between genetic divergence and time (a molecular clock) in your data. The software is freely available from *http://tree.bio.ed.ac.uk/software/tempest.*

*How does it work? TempEst performs 'root-to-tip' linear regressions, which can be used as a simple diagnostic tool for molecular clock models. It implies plotting the genetic divergence of the sequences (i.e. the sum of the branch lengths from a sequence -the tip- to the root of the tree) against the sampling time of the sequences (**Figure 4**). A linear trend with few residuals indicates that evolution follows a strict molecular clock. The same trend with greater scatter from the regression line suggests that a relaxed molecular clock model may be most appropriate. No trend at all indicates that the data contains little temporal signal and is unsuitable for inference using phylogenetic molecular clock models.*

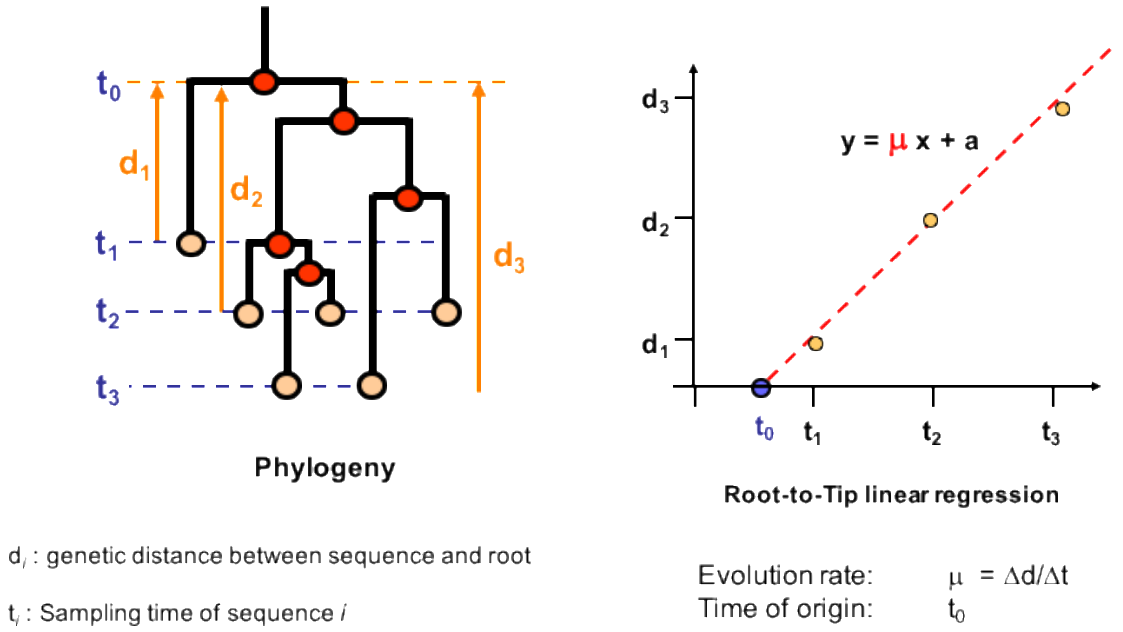


Figure 4. Phylogeny of 6 sequences (left) and corresponding root-to-tip linear regression (right)

1.2. Import the neighbor-joining tree you saved earlier or, if you can't find it anymore, select the file named '*H1N1.flu.2009.nj.tre*' in your folder. If you are not prompted for the tree when TempEst opens, import it using the drop-down menu: *File > Open…*

1.3. Tag the sequences with their sampling date.

A list of all the sequences in the tree will appear (in the default '**Sample Dates'** tab). In order to plot root-to-tip genetic distances against sampling time, each sequence has to be associated with its

date of sampling. Sampling dates are indicated at the end of the sequences' name, preceded by the suffix '_d', in the format YYYY-MM-DD. For instance, the sequence named 'A/Lima/WRAIR1687P/2009_d2009-06-27' was sampled the 27th of June 2017.

In the **Sample Date** tab, click on '**Guess Dates'**. In the '**Guess Dates for Taxa**' window, select the following options:

- The date is given by a numerical field in the taxon label that is defined by a prefix and its order (Order: **Last**; Prefix: **_d**)
- Parse as a calendar date (Date format: **yyy-MM-dd**)

1.4. Click **OK** and the time of sampling will appear in the '**Date**' column.

1.5. Go to the **Root-to-tip** tab to see the root-to-tip linear regression plot. Summary statistics of the plot are shown in the left hand side window:

- **Date range**: the maximum time interval between two sampled sequences.
- **Slope**: The slope of the regression line; corresponds to the rate of evolution, here the average number of nucleotide substitutions per unit of time (here, per year).
- **X-intercept**: the time at which the viral population had no genetic diversity, which corresponds to the time of the most recent common ancestor of the sampled population (tMRCA).
- **Correlation Coefficient**: A measure of the relationship between time and the number of accumulated mutations (range: -1, 1). A positive/negative value implies a positive/negative linear relationship between time and diversity (i.e. as time increases, so does genetic divergence). A value close to 0 suggests no relationship between time and genetic divergence.
- **R squared**: A measure of how close the data is to the regression line, i.e. what proportion of the variation in genetic divergence is explained by a strict molecular clock hypothesis.

Q3: Can we assume that the rate of evolution of the viruses in the phylogeny is constant over time (i.e. that the molecular clock is strict)?

2 – Dating migration events

Now that the properties of the molecular clock in our dataset have been established, we will estimate the time(s) at which H1N1/09 was introduced in England.

Reminder. In a 'classic' phylogenetic tree, branch lengths reflect as the number of nucleotide substitutions per site. Rates of evolution are expressed as the number of substitutions per site per unit of time. By dividing the length of a branch by the rate of evolution, we end up with a tree where branch lengths represent time units. A branch therefore represents the time elapsed between two nodes. Or, in our case, the time elapsed between two transmission events.

The reconstruction of dated phylogenetic trees is computationally intense and could not be achieved within the time imparted for this practical. A dated H1N1/09 phylogeny was therefore built prior to the session, under the appropriate molecular clock model, using the Bayesian MCMC approach implemented in the software **BEAST** v.1.8 [3]. The resulting dated tree is in your folder under the name '*H1N1.flu.2009.mol_clock.tre'.*

How does it work? Each tip of the tree has a known time, given by the sampling date of the sequence. Internal nodes are given arbitrary starting times consistent with their order in the tree (from the tips to the root). An additional parameter, the evolution rate, is used to scale these times into expected number of nucleotide substitutions per site. Markov chain Monte Carlo integration is then used to summarize the probability density function of a model tested against the data, providing a representative sample of parameter values of the chosen model. The model includes the tree topology, the times of internal nodes and the evolution rate.

We will use the program **FigTree** to display the dated tree and its annotations. FigTree is a tree editor with a graphic interface and is freely available at *http://tree.bio.ed.ac.uk/software/figtree/. It* runs on all operating systems.

2.1. Open **FigTree** by clicking on the program icon.

2.2. Import the dated tree*: File > Open… > H1N1.flu.2009.mol_clock.tre*

A phylogeny will appear. Go to Tree > Increasing Node Order to display the tree in the same way as the one you generated with MEGA. This will ease comparison. Note that the clustering pattern itself doesn't change, just the order in which the clusters are organised from top to bottom.

The dated tree should be very similar to the neighbor-joining tree you reconstructed in Session 1. However, in this tree, the branch lengths represent *days* rather than genetic distances. Notice the scale at the bottom of the tree.

2.3. On the left hand side toolbar, tick the '**Node Labels**' box. The age of the nodes will appear.

The age of a node is expressed as the number of days prior to the most recent sampling date in the tree. Here, the most recent sample is *A/Lima/WRAIR1687P/2009*, sampled on the 27th of June 2009. If, for instance, a node age equals 21 days, it means that the date at this node is the 6th of June 2009.

Q4: According to the molecular clock dating, what are the date(s) of introduction of H1N1/09 in England (rounded down)?

Tip: If mental arithmetic is not your thing, you can use the 2009 calendar provided as ***Appendix*** *to back-calculate the date of the nodes.*

Section 3. Identifying the geographical origin(s) of the H1N1/09 strains imported in England

We will finally reconstruct the migration pathways of these H1N1/09 strains.

Reminder. When individuals are infected in one location and then move to another, or infect someone whilst travelling, this is apparent as a "change" in the location ascribed to one branch of the tree. These changes in location along a phylogenetic tree can be inferred from the location values at the tips and the shape of the tree (see ***Figure 4****). To do so, a model of location exchange process is fitted to the data and the most likely location of the viral strain positioned at the nodes of the tree, together with its probability, can be estimated using a MCMC sampling procedure similar to the one used in Section 2.*

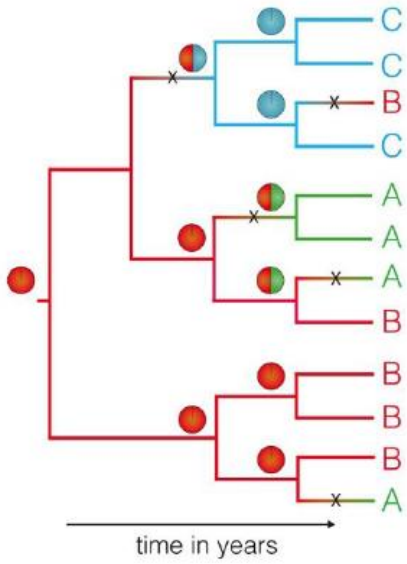


Figure 4. Principle of phylogeographic inference. Traits A, B and C at the tips of the phylogeny represent geographic locations from which genetic sequence data was collected. Crosses on the branches represent estimated changes in location. The color-coded pie charts represent posterior probability support for the location estimates. Adapted from Faria *et al.* 2014 [4].

This approach was applied to the H1N1/09 phylogeny, using an asymmetric continuous-time Markov chain [5], as implemented in the program BEAST. The asymmetric model uses separate parameters for forward and reverse rates of movement between each pair of locations. The tree file we used for the molecular clock analysis (*H1N1.flu.2009.mol_clock.tre) also* contains the result of the inferred migration patterns.

3.1. On the left hand side toolbar, pull down the **Node Labels** menu by clicking on the corresponding forward arrow (▶). Under **Display**, select '**Location'**. The most likely location of the ancestral virus located at the nodes will appear.

3.2. To display the posterior probability of the most likely node location, select '**Location.prob**' in the **Display** menu of the **Node Labels**.

Q5: Which country is the most likely source of the H1N1/09 epidemic? What is the probability of that location?

Q6: Where were the English strains of H1N1/09 imported from? How confident are we? How does your answer compare to that of Q4?

References

[1] N. Saitou and M. Nei, 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, Jul. 1987.
[2] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus, 'Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)', *Virus Evol.*, vol. 2, no. 1, p. vew007, Jan. 2016.
[3] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Mol. Biol. Evol.*, vol. 29, no. 8, pp. 1969–1973, Jan. 2012.
[4] N. R. Faria *et al.*, 'The early spread and epidemic ignition of HIV-1 in human populations', *Science*, vol. 346, no. 6205, pp. 56–61, Mar. 2014.
[5] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, 'Bayesian phylogeography finds its roots', *PLoS Comput. Biol.*, vol. 5, no. 9, p. e1000520, Sep. 2009.

Appendix – 2009 Calendar

JANUARY

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

FEBRUARY

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28

MARCH

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

APRIL

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

MAY

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

JUNE

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

JULY

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

AUGUST

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

SEPTEMBER

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

OCTOBER

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

NOVEMBER

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

DECEMBER

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Introduction to Infectious Disease Modelling and its Applications – 2018

Session 22: Applications of stochastic models: Estimating R_n for eliminated and emerging diseases

Practical

Overview and Objectives

In this practical you will use a stochastic model of infection transmission as a tool to study the epidemiology of a communicable disease after elimination

The motivating example is measles, where vaccination programmes have led to elimination in many countries. Despite elimination, clusters of cases (seeded by importations from countries where measles hasn't been eliminated) continue to occur. In this practical you will use the same stochastic model in two complementary ways: predictively, to tell you how likely different outbreak sizes are for given values of the net (or effective) reproduction number, R_n (also sometimes referred to as R_t or, simply R); and inferentially, to infer the likelihood of different values of R_n for given outbreak sizes.

The approach presented can also be used to investigate the epidemiology of a new potentially emerging human pathogen capable of sporadic, but not yet sustained, human-to-human spread. The common thread connecting these two applications is that in both cases, the net reproduction number, R_n, is less than one. For measles this is because vaccination has removed most susceptibles (remember that $R_n = R_0 \times$ proportion of population who are susceptible). For a potentially emerging infectious disease it is because the pathogen has yet to evolve to become capable of sustained human-to-human transmission (so everyone is susceptible, but $R_0<1$).

In both cases, by appropriate analysis of the frequency distribution of the size of observed clusters of cases we can make inferences about the value of R_n , allowing us to assess how close we are to control failure (R_n >1). This has obvious public health implications: in the case of measles if we cannot be confident that R_n is below one, we would probably want to vaccinate more people in order to eliminate the risk of a major epidemic. In the case of a new pathogen, such as a hyper-pathogenic avian influenza strain, as R_n gets closer to one we might want to step up measures to control human-to-human spread and to prepare containment measures to increase the chance of controlling an outbreak at source.

By the end of this session you should:

- Have an appreciation of the importance of stochasticity in determining the size of clusters/outbreaks when R_n is less than one.
- Understand how the distribution of sizes of clusters /outbreaks is determined by the initial net reproduction number.
- Understand how cluster size data can be used to make inferences about R_n using a maximum likelihood approach.
- Appreciate the important public health implications of such analysis.

Background

You should remember from the lectures that when R_n >1, introducing one infectious case into a population otherwise free of infection can, with some probability, produce a major outbreak. There is also a chance that the epidemic will not take off, but will instead *fade out* (come to an end) after only a small number of secondary cases.

In contrast, when R_n<1 introducing just one case will never produce a major outbreak. Transmission will always be self-limiting, with all outbreaks coming to an end on their own. In a large population, only a very small proportion of people will be infected. Nonetheless, if 0< R_n<1 there is still a chance of some secondary transmission and, although there will not be a major outbreak, relatively small outbreaks or clusters of cases will occur.

This practical uses the approach described in reference [1] to monitor the status of elimination (reference [2] gives an example of an application of the same approach for assessing the risk from a potentially emerging pathogen).

In this practical you are first asked to simulate outbreak data for different R_n values. This should give you an insight into how chance and the value of R_n combine to determine the outbreak size. Then you are asked to use distributions of outbreak sizes to estimate R_n using real measles data from England and Wales.

You should aim to spend no more than 45 minutes on part I, before moving on to part II. The optional questions are there for anyone who wants a deeper understanding of the likelihood method used in part II..

PART I: Simulating outbreaks when $R_n \leq 1$

1. Open up the spreadsheet *branching process models.xls*. If there is a security warning below the ribbon stating “Macros have been disabled”, click on the “Options” button next to this warning and then select the “Enable this content” option, before clicking on OK. You will need macros to be able to run the practical.

This spreadsheet is designed to simulate outbreaks using method 2. At the top of the spreadsheet in the *simulations* workbook (under the heading “Simulation set 1”) you should see:

a) **Pink cells** (E3, B10) containing numbers you can edit to change the number of simulations to run and the R_n value.
b) **Green cells** (row 11) containing the number of cases in each generation of the outbreak in the most recent simulation run (these are replaced by grey cells when the outbreak in the current simulation has terminated)
c) A **green chart** (row 13) plotting the number of cases in each generation for the most recent simulation run.
d) A **yellow cell** (CZ11) showing the total number of infected cases (including generations 1 to 100) in the most recent simulation run.
e) **Two buttons**. The grey button allows you to *Run simulations* (the number of simulations to run is specified in cell E3). The yellow “show histogram” button brings up a table and graph that summarises the simulation results (don’t click on this button just yet).

Note that by default columns R to CV are hidden (representing generations 14 to 97 of the simulation). This is because it's not possible to display all 100 generations on the screen at one time. Feel free to unhide them if you like.

The simulations are set up with a population of 100,000 and start with one initial case in the first generation (cell E11). This case will infect a random number of secondary cases. Each person in each generation has a probability of R_n/100,000 of being infected by any particular case. The mean number infected by one case is therefore R_n (100,000, times R_n/100,000) the current value of the net reproduction number (cell B10).

Q1.0 Look at the formulae in cell G11 (which is copied in the 99 cells to the right of E11). See the Appendix if this is the first time that you have seen Excel's critbinom function.

Q1.0 Can you explain how this formula works?

Since we have a fixed number of trials (people) each with the same probability of success (infection) the distribution of the number of infected cases is binomial[1]. The cases infected by the first case are the second generation (cell F11), and each of those can also infect a random number of cases, with the same mean (R_n) and probability distribution. Those infected by the second generation are the third generation (cell G11) and so on. The outbreak comes to an end when a generation has zero cases (shown in grey).

This type of model is called a *branching process* model because it gives rise to a tree structure: the initial case is the trunk, second generation cases are branches coming out of the trunk, and so on (see figure 1). The key difference from a stochastic epidemic model is that the size of the susceptible population is assumed not to change. This is a reasonable assumption when R_n <1, as only a very small proportion of the total population will be affected (assuming the total population is large). This branching process model would therefore not be appropriate when R_n>1 because in that case a major epidemic could occur. When a major epidemic does occur it is the decline in the susceptible population that causes the epidemic to come to an end.

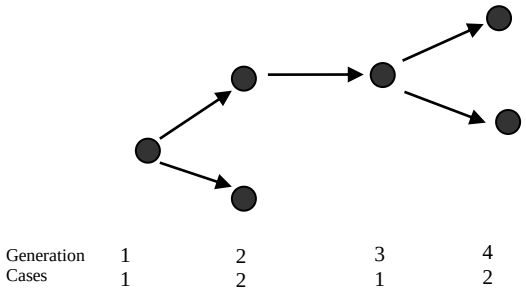


Figure 1 Schematic illustration of a branching process

Q1.1 Before making any simulation runs try to predict what output you expect the model to give. What do you expect to happen when you run simulations with i) R_n =0.99 ii) R_n =0.5 iii) R_n =0 .01

2. Test these hypotheses. First set the R_n value in cell B10. Then set the number of simulations to run in cell E3 to 1, and click on the "Run simulations" button. Every time you click on this button a new simulation run will be performed. For each value of R_n try performing about 10 to 20 simulations.

Q1.2 How important is chance in determining what happens for each of the three R_n values?

The next step is to collect the results from the individual simulation runs, and use these results to estimate the probabilities of different outbreak sizes for different R_n values. Click on the yellow "Show histogram" button. A yellow table listing the frequency of different outbreak sizes from all simulation runs (cells F15:G26) and next to it a histogram plotting the same information should appear.

There are also two new buttons. The *Reset* button deletes all previous simulations runs in the current simulation set, and resets the yellow table and histogram. The *Refresh* button just updates the histogram and table to include the most recent simulation run.

3.Now click on the *Reset* button to clear the histogram. Set R_n to a value of 0.5 in cell B10, and set the number of simulations to run to 20 (cell E3). Click the *Run simulations* button to run 20 simulations. Note what the yellow histogram looks like. Perform another 200 simulation runs in batches of 20 and note how the histogram changes

You should notice that the overall shape of the histogram changes considerably as you perform more simulation runs.

Q1.3 What do you expect to happen to the shape of the histogram as you perform more and more runs? Test your hypothesis by performing another 500 simulations (enter 500 in cell E3 and click on Run simulations).

4. Repeat step 3 and question 1.3 with an R_n value of 0.99 (you will first need to reset the histogram by clicking on the *Reset (delete all runs and start again)* button).

Q1.4 What do you expect to happen in the simulations when R_n is set to 1.1? Why? What do you expect to happen to the frequency distribution represented by the histogram?

5. Test these predictions: first click the *Reset* button to clear the previous histograms and set R_n to 1.1 and cell E3 to 1 (i.e. just run one simulation at a time). Then perform several individual simulation runs, looking at the course both of the individual simulations (the green graph) and the frequency distribution of the outbreak size.

Sometimes you might notice that the histogram goes blank and frequencies are listed as "#NUM". This happens because Excel has encountered a number that is too large for it to handle. When this happens simply click on the "Reset" button and carry on.

6. So far you have explored the behaviour of a simple branching process model and (hopefully) tried to relate it to the epidemic theory you have learned. The next step is to use simulations more systematically to estimate the probabilities of different size outbreaks for different R_n values.

6.1 Clear the previous runs by clicking on the reset button. Set R_n to 0.2, and perform 1000 simulation runs. This represents simulation set 1.

.

6.2 Scroll down the page to "simulation set 2" (rows 29-57). You should see a second copy of the controls you have been using (although without the green graph). In this space, set R_n to 0.4 (in cell B38). Clear any pre-existing histogram, and perform 1000 simulation runs with this R_n value. Then perform 1000 runs for an R_n value of 0.6 under "simulation set 3" (rows 58-86), 1000 runs for an R_n value of 0.8 under "simulation set 4" (rows 87-115), and 1000 runs for an R_n value of 1.0 under "simulation set 5" (rows 116-144)

6-4 If you now scroll down again, you should see the results of all the simulations you have just performed in rows 146-195. The histograms you have created have been duplicated and the probabilities of outbreaks of different sizes as estimated from your simulations have been tabulated and plotted in a single stacked bar chart (on the right hand side, inside the grey box – you may need to click on its *Refresh* button to update it to include the latest simulation). Both the histograms and the stacked bar chart show the same information: the probability of an outbreak being of a particular size for a given value of R_n.

We can use the information in the stacked bar chart in two ways. If we know what the R_n value is, we can ask how likely it is that an outbreak will be of a particular size. Alternatively, if we don't know R_n, but want to estimate it from data on the size of a particular outbreak, we can estimate it by asking which value of R_n is most likely for the given outbreak size. In the cases of measles and avian influenza this is exactly the kind of question we are interested in.

Q1.5 Suppose that there is an outbreak of an infectious disease and seven people are infected. No attempt to control the outbreak is made, and the outbreak comes to an end of its own accord. Suppose also that you are sure that R_n is less than one, but would like to know what it is more exactly. Based on the values in your stacked bar chart, which of the five possible values for R_n is most likely?

PART II: Estimating R_n from outbreak data

The most likely value of R_n for some particular outbreak data is called the *maximum likelihood estimate* (or MLE). Your answer to Q 1.5 was your MLE for R_n given the data (one outbreak affecting seven people). The MLE is just the value of the parameter that would be associated with the highest chance of observing the particular data (assuming the model is true). The process of finding the MLE is called *maximum likelihood estimation* and is one of the most widely used methods for estimating parameter values (it was also used in the practical where you analysed seroprevalence data). This second part of the practical shows how maximum likelihood estimation can be used to estimate R_n in more complex situations when there are data from multiple outbreaks. By making inferences about R_n in this way, we can find how close we are to control failure (R_n >1) and take precautionary measures as necessary.

1. Select the worksheet *Inference* in the *branching process models.xls* spreadsheet by clicking on its tab at the bottom of the screen.

In the top left corner you should see a blue table showing the "Probability of specified outbreak size for different reproduction numbers" and next to it a stacked bar chart that looks very similar to the one you have just created. Apart from having more R_n values, the only difference is that in this one the values are calculated *analytically* (i.e. using an exact formula) rather than by simulation, as in the version you created[2]. You should find your simulated estimates for the probabilities are very similar to the exact values given here. By performing more simulations your estimates could be made arbitrarily close to the exact values.

Q2.1 Repeat question 1.5 but now using the blue table in rows 6-19 (or the stacked bar chart derived from it) in the inference worksheet. According to this table (or chart), what is the MLE for R_n if there is one outbreak with 7 cases?

Real measles data from England and Wales appears in the yellow table (rows 24-33). These come from the four years following a national measles vaccination campaign in November 1994 aimed at all children aged 5-16 years. The turquoise table below that (rows 38 to 52) is used for calculating the likelihood of different R_n values. To use it you enter the number of outbreaks of each size in the pink cells (E39:K39). So, for example, if there was just one outbreak of seven cases, you would enter a one in cell H39 and a zero elsewhere. Below this table there is a graph plotting the likelihood calculated from this table against R_n.

Q2.2 Now use the turquoise table for calculating likelihoods (rows 38:52). First check that you get the same answer to question 2.1 when you use the pink cells to specify that there was one outbreak with 7 cases. Then use the pink cells to specify that there was one outbreak of 2 cases. Which value of R_n gives the maximum likelihood in this case? You can check that your answer agrees with the numbers in the blue table (& stacked bar chart). What is the likelihood at this value?

Note: likelihoods for different R_n values are shown in column M of the turquoise table, and the corresponding logarithms of these values (log likelihoods) are shown in column O, with the largest value highlighted in red. Log likelihoods are usually preferred to likelihoods because likelihoods often take such small values that they become hard for computers (and people) to handle.

Q2.3 Use the pink cells (E41:K41) to enter data on the size of measles outbreaks for England & Wales. Do this first for the 1995 outbreaks, and find the MLE for R_n for this year (taking into account all three outbreaks). Repeat, obtaining MLEs for the other three years. As you do this, note how the graph below the table showing the likelihood of different R_n values changes shape.

One problem with the above approach is that large outbreaks may be more likely to be reported than small outbreaks. In the case of the England & Wales data outbreaks of size one and two do not appear in the yellow table at all. i.e. we can assume that there is no chance of outbreaks of size one or two being reported. Failing to account for the fact that outbreaks of size one or two have not been reported would lead to biased estimates of R_n.

Q2.4 Would you expect the estimates of R_n obtained without accounting for the non-reporting of outbreaks of size one or two to be too high (biased upwards) or too low (biased downwards)?

You can check your answer to Q2.4 using the table for calculating likelihoods in rows 103-117. This has modified the previous calculations by accounting for the fact that outbreaks of size 1 and 2 are not reported: the numbers in the blue table (rows 75-89) have been modified and are now the probabilities of specified outbreak sizes given that outbreaks are only reported when at least three people have been effected (i.e. outbreaks affecting fewer than three people are discarded as the reporting of these is assumed to be unreliable).

Q2.5 Recalculate the MLEs for R_n for each of the four years of the UK data accounting for the fact that outbreaks of size one or two are not reported using the table in rows 103-117. Do your answers agree with you predictions from question 2.4?

You may also have noticed one apparent limitation of the approach for estimating R_n from outbreak data presented here: it only seems to apply to situations where you know that R_n <1. What if you can't be so sure of this fact? In fact, the method can be extended to allow for the possibility of R_n values greater than one by considering outbreaks above a certain size to be censored. The technical details of this approach are described in Farrington *et al*. (2003).

The basic idea is that if R_n >1 there is a chance that outbreaks will not be limited to just a few cases but will grow exponentially (initially), affecting many thousands. In such cases the branching process model breaks down: while the branching process would continue to grow exponentially indefinitely, in reality the epidemic comes to an end because susceptibles get used up. However, by considering outbreaks above a certain size (1000 cases, say) to be censored, we can still use the branching process model to estimate R_n from outbreak size data, even when R_n may be greater than one. This approach has been implemented in the spreadsheet in rows 140-230

Q2.6 Use this approach (& the table in rows 185-209) to obtain a single MLE for the R_n for measles in England & Wales over the entire period 1995-1998. The one 1997 outbreak listed as affecting 100+ cases should be classified as 100-999.

You might have noticed that for years with more data the turquoise likelihood curve gets "pointier", ie. the peaks get sharper. When there is no data the graph is completely flat. This suggests that the "pointiness" of the peak can be used to quantify the uncertainty in the MLEs: the sharper the peak, the more certain the estimate. This does indeed turn out to be the case, and this idea can be made more precise. In fact, we can construct approximate 95% confidence intervals about our MLE for R_n by finding the values of R_n for which the log likelihood is 1.92 lower than that of the MLE.

Q2.7 Use the above fact to estimate approximate 95% CIs for the R_n for measles in England & Wales over the entire period 1995-1998. Use the table in rows 185-209 and don't worry about being too precise: you will only be able to do this very roughly to an accuracy of one decimal place. Based on your answer, how confident would you be that the measles vaccination program had been successful at reducing R_n to below one?

Q2.8 How would your conclusions be affected if the outbreak in 1997 affecting "100+" cases had, in fact, affected ">999" cases? In this case, what action would you recommend?

If you have time try the optional questions below. You may also want to try the supplementary questions for this practical, which illustrate how you can set up models using method 3 described in the lecture notes.

Optional questions

By looking at the contents of cell M45, and then the contents of the cells M45 refers to you should be able to confirm that the likelihood is calculated using the following formula

[probability of an outbreak of size 1 given R_n is 0.5]$^{\text{number of outbreaks of size 1}}$
×
[probability of an outbreak of size 2 given R_n is 0.5] $^{\text{number of outbreaks of size 2}}$
×
[probability of an outbreak of size 3-4 given R_n is 0.5] $^{\text{number of outbreaks of size 3-4}}$
× ….

Note that Excel uses the symbol "^" to mean *raised to the power of*..

To see where this formula comes from remember that if the probability of thing A happening is p, and the probability of thing B happening is q, then, provided A and B are independent, the probability of A and B both happening is $p \times q$. The above formula simply assumes that the outbreaks are independent.

Q2.9 If $p_{2|R=0.5}$ is the probability of an outbreak of size 2 given that R_n is 0.5, what is the probability that if two independent outbreaks are observed both have size 2 given that R_n is 0.5? Give your answer as both a formula (in terms of $p_{2|R=0.5}$) and as a number (taking the value for $p_{2|R=0.5}$ from the table in rows 6-19). Confirm this by entering 2 in cell F39, and checking the value of the likelihood when R_n is 0.5.

Q2.10 Try to write down an algebraic expression for the combined likelihood that there are 7 outbreaks of size 3-4, and 3 outbreaks of size 5-9 , given that 10 independent outbreaks occur and that R_n =0.8. Use the same notation as in question 2.5, so, for example $p_{3\text{-}4|R=0.8}$ is the probability of an outbreak of size 3-4 given that R_n is 0.8.

Q2.11

If we want more precise estimates of R_n (i.e. more decimal points) we could use Excel's Solver (in the tools menu) to find the precise value of R_n that maximises the likelihood. The formula required for the likelihood is given in the worksheet "analytic results". Mathematically inclined students might like to implement this Solver-based maximum likelihood estimation, and calculate the MLE for a single outbreaks of sizes 2, 3, 4 and 5, and then conjecture a relationship between the MLE and the outbreak size.

Notes

1. The analytical results in part 2 of the practical in fact assume the secondary cases have Poisson rather than a binomial distribution. However, when n is large and p is small the binomial distribution provides an excellent approximation to the Poisson distribution. Because Excel has no efficient way to simulate Poisson random numbers, a binomial distribution is used instead here. Note also that the actual population size is arbitrary. The results would hardly be affected if we changed the population size from 100,000 to

100,000,000. This is because we are concerned here only with small outbreaks, where the depletion of susceptibles can be ignored.

2. In this case the exact solutions can be calculated because our underlying branching process model is simple enough to be fully analysed mathematically. When this is done the outbreak sizes can be shown to come from something called the Borel-Tanner distribution (you can see the underlying calculations in the worksheet called “analytic results”). In general, however, such analytic solutions aren’t possible for stochastic epidemic models, and we need to use simulation instead.

References

[1] De Serres G, Gay NJ, Farrington CP (2000). Epidemiology of transmissible diseases after elimination. Am J Epidemiol. 151:1039-48.

[2] Ferguson NM, Fraser C, Donnelly CA, Ghani AC, Anderson RM.
Science. (2004); 304:968-9. Public health risk from the avian H5N1 influenza epidemic.

[3] Farrington CP, Kanaan MN, Gay NJ (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 4, 279-295.

Appendix – Reminder of the Reed-Frost equation

As you saw in the lecture, the Reed-Frost model gives the following equation for the probability that exactly k out of S_t susceptibles at time t are infected and develop disease by time t+1:

$$P(I_{t+1} = k) = \binom{S_t}{k} \lambda_t^k (1 - \lambda_t)^{S_t - k}$$

λ_t is here defined as the risk that a susceptible individual is infected by at least one case between time t and t+1.

Excel has a function called “critbinom” which works as the inverse of the binomial distribution i.e. it identifies the number of cases which will be seen for given values for the cumulative probability, the number of susceptibles at time t (S_t) and λ_t . The notation is as follows:

=CRITBINOM(S_t, λ_t , cumulative_prob)