EEB 200A: Approximate Bayesian Computation (as easy as ABC…)

Dr. Kirk E. Lohmueller

November 23, 2015

Estimation of population size

Here you will implement an ABC approach to estimate the effective population size.

Imagine that you resequenced the same 100kb region of two Y chromosomes. The Y chromosomes are non-recombining. You observe 50 SNPs within this region.

You would like to use this (albeit very limited) dataset to estimate the effective population size of males.

Based on other information, you know from a very reliable source that the mutation rate per base pair on the Y chromosome is $\mu=1 \times 10^{-8}$ per base pair per chromosome per generation.

Because we are considering a non-recombining region, this problem can be easily addressed using the coalescent without recombination for a sample size of $n=2$. (hint, this is what we spent a lot of time in class discussing).

Let's go ahead and develop an ABC approach to do this.

1) Implement the following ABC rejection sampling algorithm in R:

> a) Assume that N can be any value between 100 and 100,000. Draw 1 million values from the prior distribution of N. Let's use a uniform distribution for N. Assume that $N \sim U[100,100000]$.
>
> b) For each value of N, simulate a TMRCA. Note, you should use the same approach as on last week's assignment (i.e. Draw times from an exponential distribution). Second hint: The total number of Y chromosomes in the population is N, rather than $2N$ (because the Y chromosome is haploid). Adjust your rates of coalescence accordingly.
>
> c) For each TMRCA, add a Poisson number of mutations with the appropriate mutation rate. Again, this will follow from what you did last week.
>
> d) We now need to decide which of the million draws from the prior give data that are "close" to the observed number of SNPs in the actual data and should be

accepted. To do this, let's accept all values of *N* that give somewhere between 45-55 SNPs.

e) Congratulations! You should be done!

2) Make a density plot of your prior distribution and your posterior distribution of *N*. Please plot them on the same axes and be sure to label which line corresponds to the prior and which corresponds to the posterior.

3) What is the median value of the posterior distribution of *N*?

4) Generate a 95% credible interval for the posterior distribution of *N* (like a confidence interval, but for Bayesians. Note, in this framework, the there actually is a 95% chance of the true parameter value falling in this region. This is not the case for normal frequentist confidence intervals). Hint, use the "quantile" function in R.

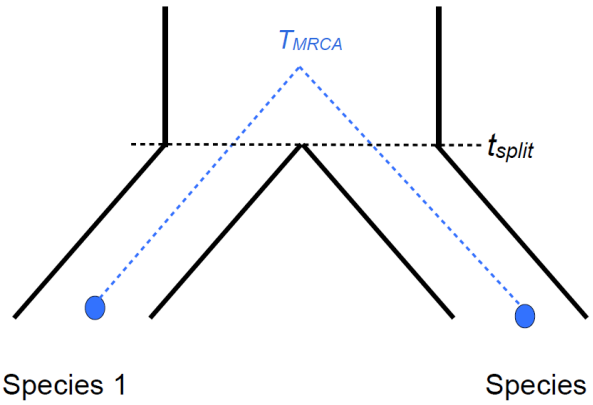
5) How does the posterior distribution differ from the prior distribution? A descriptive answer here will suffice. The degree to which the posterior differs from the prior relates to the amount of information in the data.

6) Repeat your ABC analysis, but change the prior distribution of *N* to be U~[1000,1000000]. What is the mean, median, and 95% credible interval for the posterior distribution. How does this differ from what you computed in questions 3-4 for the original prior distribution? What does this tell you about the effect of the prior distribution for in Bayesian statistics?

7) If you wanted a more precise estimate of the population size (assume that there is no sex biased demographic history so that you can easily extrapolate the total population size from the Y chromosome population size and vice versa), would you be better off: A) sequencing a bigger region of the Y chromosome, or B) sequencing the same amount on the autosomes? Why?

Estimation of population split times

Imagine the following model:



You have sampled 1 Y chromosome from each of two species. You have sequenced the same 100kb region as in the question above. Now, you find 600 SNPs or differences between the pair of sequences. Assume that no migration or gene flow has occurred since the split.

As before, you know from a very reliable source that the mutation rate per base pair on the Y chromosome is $\mu=1 \times 10^{-8}$ per copy per generation.

Initially, assume that the ancestral population size has N=100,000. Again, because we're analyzing the Y-chromosome, everything is haploid and in terms of numbers of chromosomes.

Your goal is to estimate t_{split} from the number of differences between the two sequences using an ABC approach.

Before we begin, let's formulate the model.

1) Once the two chromosomes make it back into the ancestral population (of size N=100,000), what is the expected amount of additional time we have to wait until they coalesce?

2) What is the expected time until the two chromosomes coalesce with each other? You should write this formula (use t_{split} as the split time, rather than a specific number, because you haven't estimated this number yet!). Hint: Break the total time into 2 parts: the time from the present day till t_{split} and the time for what happens in the ancestral population (i.e. your answer for part 1) of this question).

3) Implement the following ABC approach in R to estimate t_{split}:

> a) Assume that t_{split} can be any value between 50,000 and 1,000,000 generations. Draw 1 million values from the prior distribution of t_{split}. Assume that t_{split} ~U[1000,1000000].
>
> b) For each value of t_{split}, simulate a TMRCA. Note, you should use the same approach as on the first part of this assignment. But, keep in mind that the

coalescent time is a function of both the split time (t_{split}, which is drawn from your prior) and the coalescent time in the ancestral population.

c) For each TMRCA, add a Poisson number of mutations with the appropriate mutation rate. Again, this will follow from what you did last week.

d) We now need to decide which of the million draws from the prior generate data that are "close" to the observed number of SNPs in the actual data and should be accepted. To do this, let's accept all values of t_{split} that give somewhere between 550 and 650 SNPs.

e) Congratulations! You should be done!

4) Make a density plot of your prior distribution and your posterior distribution of t_{split}. Again, plot both the prior and the posterior on the same axes, and label which is which.

5) What is the median value of the posterior distribution of t_{split}?

6) Generate a 95% credible interval for the posterior distribution of t_{split}. Hint, use the "quantile" function in R.

7) How does the posterior distribution differ from the prior distribution? A descriptive answer here will suffice. The degree to which the posterior differs from the prior relates to the amount of information in the data.

8) What if we did not know the true value of N? Repeat the ABC approach to estimate t_{split}, but this time, rather than assuming that N=100,000, draw N from a N~U[1000,1000000] distribution.

9) Do your estimates of the median and credible interval of t_{split} differ from above?

10) Did you have fun with the YM[R]CA?