IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

EXAM 1

For Internal Students of Imperial College of Science, Technology and Medicine

Exam Date: Friday, 09th Jan 2017, 1000 – 1300

Length of Exam: 3 HOURS

Instructions:

Please note that this exam has three Sections:

- SECTION 1 requires ONE of two questions to be answered
- SECTION 2 requires TWO of three questions to be answered
- SECTION 3 requires ONE of two questions to be answered

THUS, A TOTAL OF FOUR QUESTIONS ARE TO BE ANSWERED, EACH CARRYING EQUAL WEIGHTAGE (25 pts each). So it is a reasonable guideline to spend about 45 minutes on each question.

Read the instructions carefully at the head of each section.

PLEASE PUT ANSWERS TO EACH QUESTION IN A SEPARATE EXAM BOOK.

WE REALLY MEAN IT. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

Continues on next page

Computing, Statistics, Model Fitting

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. You have obtained some experimental data called `some_data.csv` consisting of three variables x, y, z. You are planning analyze these data, with the objective of examining the relationship between x, y & z.

(i) State the principles of a reproducible analysis workflow, and explainhow would you set it up for this dataset. Write the R/Python (or pseudo-code) commands, and add a single-line comment to state what each one does (30%) as you would do in the actual script.

(ii) How would you import and explore these data once the workflow is set up? Please write out the appropriate R/Python (or pseudo-code) code in neat, commented blocks of code fragments/commands, and explain why you would use each command with a single-line comment, as you would in the actual script. (70%)

B. Please answer the following (each question equally weighted):

(i) Explain type 1 and type 2 errors, and how we deal with those when we conduct data analysis.

(ii) Elaborate the consequences for scientific progress of not considering errors, and how these consequences differ for both type of errors.

Continues on next page

GIS, Genomics, Population Genetics

Please select exactly **two questions** and answer them. Please indicate clearly each answer book which question you are answering.

A. Land use and land cover change (LULCC) models could be an important tool for understanding present day, and predicting future, patterns of biodiversity. Describe the level of certainty that we can place in LULCC model predictions, and either defend or attack the assertion that uncertainty in LULCC models does not adversely affect the confidence with which we can extrapolate biodiversity patterns.

B. The Breeders Equation, $R = Sh^2$ is commonly used by breeders of domestic animals, however, evolutionary biologists also have made use of it.

Define and explain each variable of the Breeder's equation in detail. Explain a good approach to quantify each variable in wild populations. Highlight problems and pitfalls.

C. Inferring demographic history from genetic data with confidence requires sufficient statistical power. Why are large samples necessary to distinguish between different demographic scenarios? What are the relative roles of numbers of genetic loci and numbers of sampled individuals with respect to the necessary sample sizes for inferring demography?

Continues on next page

Neutral theory, modelling, model fitting, HPC

Please select exactly **one question** and answer it. Please indicate clearly in your answer book which question you are answering.

A. Answer the following questions. Please be brief in your answers to these questions – bullet points are OK as you are being marked on content:

(i) Take an individual based model of an ecological community containing several species. In this model individuals are arranged in a spatially explicit manner and may reproduce or die in each time step of the model. For each of the following scenarios would the model be considered neutral or not? Give a brief reason for each answer.

a. Individuals are chosen at random to die according to a uniform distribution; each dead individual is replaced by the offspring of another individual also chosen according to a uniform distribution. (10%)

b. Individuals are chosen at random to die according to a uniform distribution; the probability of reproduction for any individual depends on how many conspecific individuals are close by to it.(10%)

c. There are two different kinds of habitat in the model (call them habitat A and habitat B). Individuals (of any species) that happen to be growing in habit B have a greater probability of death due to increased frequency of natural disasters in that area.(10%)

d. There are two different kinds of habitat in the model (call them habitat A and habitat B). Individuals belonging to some species are more likely to die in habitat A than in habitat B, individuals belonging to other species are equally likely to die in any habitat.(10%)

(ii) You wish to conduct a set of 20 neutral simulations on a high performance computing facility for a very large habitat size. The simulations run for a burn in period until they reach dynamic equilibrium and then begin outputting data. Here is a sample from the shell script you plan to use:

```
#PBS -l walltime=2:00:00
#PBS -l select=1:ncpus=1:mem=800mb
```

a. What is meant by dynamic equilibrium and why is it necessary to have a burn in period for your simulations? (20%)

b. After submitting your test jobs to the cluster you find that they quickly fail and return no results. What most likely happened and how could you fix the problem? (20%)

c. Now you find that your simulations run for a couple of hours and return an empty file, but the file contains no data. What most likely happened and how could you fix the problem? (20%)

B. Consider the logistic map $y_{n+1} = r \times y_n \times (1 - y_n)$ which can be studied with the following pseudo code, where the parameter r is decided before the code is run and the parameter n is a very big number also decided before the code is run.

```
y <- 0.5
For (all integer values of i from 1 to n) {
    y <- r * y *(1 - y)
}
For (all integer values of i from 1 to 1000) {
    y <- r * y *(1 - y)
    Save the value of y in an output
}
```

Continues on next page

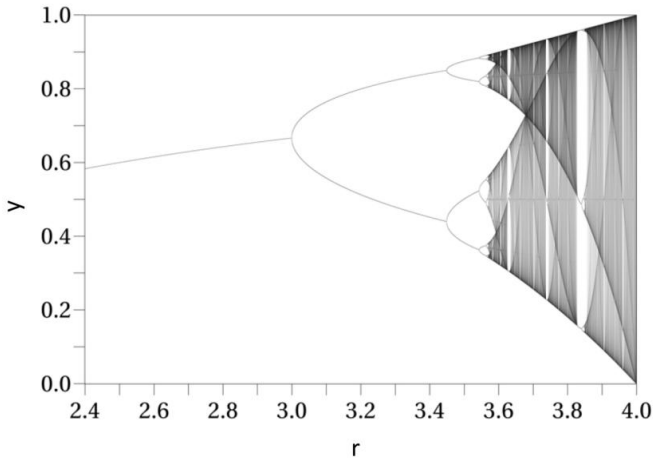
(i) Suppose $r = 2$, what value or values will get saved in the output. Explain your workings. (20 %)

(ii) Suppose r = 0.5, what value or values will get saved in the output. Explain your workings. (20 %)

(iii) What can you say about the values that might saved in the output if r = 8. Explain your workings. (20 %)

(iv) Describe what a pseudo random number generator does and what role does the random seed play in this. (20 %)

(v) If the code above is run for a wide range of different values of r, and all the corresponding saved values of y are plotted as a function of r on a graph, a portion of the result is shown below. Describe what is happening as the value of r increases. Say for which values of r might you be able to use the results as a sequence of pseudo random numbers? (20 %)



End of paper