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%)

Model Answer (Markers – Samraat Pawar (first), Julia Schroeder (second)):

Answers:

(i) A reproducible analysis workflow should allow any user to run the analysis script(s) on any computer without any manual modification to the code and data. A neat workflow would separate the analysis into `Code`, `Data`, and `Results` directories. The steps would be first `setwd()` to the analysis project directory location and then create the directories using `dir.create()`, so something like:

```
setwd("MyPathToProject/ProjectDirectory") #set working directory

dir() # check what's already there

dir.create("Code") #Create directories

dir.create("Data")

dir.create("Results")
```

Students that discuss the problem with putting a `setwd()` command at the start of the script, and show how to do the directory creation in one go, would have demonstrated a high level of expertise.

(ii) A standard data exploration would require something like (a complete answer would provide the following steps and code structure):

```
##########Import Data ##########
MyData <- read.csv("../Data/some_data.csv") #import data using a relative path
ls()#Check that the data frame has appeared

##########Explore Data distributions ##########

str(MyData) # Examine structure of data frame
head(MyData) # Examine top few rows
summary(MyData) # Examine data summary
hist(x) # Examine data distributions
hist(y)
hist(z)

##########Explore pairwise relationships##########
plot(x,y)
plot(y,z)
```

Continues on next page

```
plot(x,z)

# or better still:
pairs(MyData[1:3])

##########Save Results##########

pdf("../Results/MyHist_x.pdf")
hist(x)
dev.off()

etc.

##########The End##########
```

The answer need not be exactly this in terms of the plotting commands and code sections, but the student should cover these components of the data exploration, and explain why (s)he went though the steps, including key points like the use of a relative paths to load and save outputs. Specifically, I would expect:

- The data loading step
- Data examination steps (str(), head(), summary(), etc.)
- Data plotting to examine distributions and outliers
- Saving the results in an appropriate directory in an appropriate format

Students that cover things like dealing with NA's and explain format of the output (why pdf and not png, vector vs. raster), etc would have demonstrated a higher level of R expertise and understanding of the principles.

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.

Model Answer (Marker – Julia Schroeder (1st), Samraat Pawar (2nd)):

(i) A top answer will clearly lay out the sources for both error types, and how we assess them. It will explain p=0.05, and statistical power analysis. It will highlight the importance of sample size. A complete answer will also mention the square law of sample size and precision.

(ii) A complete answer will then discuss the impact this has on a larger scale, that 1 in 20 studies might be wrong because of p=0.05 etc.

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.

Model Answer (Marker – Rob Ewers (1st), Cris Banks-Leite (2nd)):

LULCC models were shown in the lecture to be good at predicting large-scale patterns of land use change over 10-30 year time periods, but were poor at predicting fine-scale patterns of land use change over small time periods. Lecture eg is Rosa et al 2013.

Very few models quantify uncertainty in their predictions or attempt to validate their models, but the ones that do tend to have low accuracy (e.g <5% perfect match between prediction and observation). Distance-based validations over periods of one decade, that allow for near vs far misses in predictions, can have >80% accuracy. Lecture e.g. is Rosa et al 2013.

All models assume stationarity of processes which is easily falsified. Changes to the processes driving land use change will void the predictions of future change. Lecture eg is Rosa et al 2015.

Uncertainty in LULCC models doesn't weaken our confidence for modelling large-scale extinction scenarios based on species-area relationships, because LULCC predictions at those spatial and temporal scales are robust. Lecture egs were Rosa et al 2016 and Wearn et al 2012.

LULCC modelling at fine scale, however, used for modelling terragenies and species distributions in landscapes, is not reliable and inferences about biodiversity patterns emerging from these simulations should have low confidence placed in them. Lecture eg was Ewers et al 2013.

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.

Model Answer (Markers – Julia Schroeder (1st), Jason Hodgson (2nd)):

A good answer defines each term (response to selection, selection differential and heritability) in the biological and mathematical way. Empirical quantification requires the knowledge of ancestry, fitness, and changes over time in phenotype and genotype. Pitfalls are the phenotypic gambit, and also that heritability is not a measure of individuals, but a population measure. A distinction answer explains why the Robertson Price identity, using G-matrices, is a better approach. A distinction answer can also point out that heritability decreases when the genetic variance is depleted under selection.

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?

Model Answer (Markers – Jason Hodgson (1st), Austin Burt (2nd)):

- Demographic history includes knowledge about population size and changes in population size through time.
- The amount of genetic variation in a population at mutation-drift equilibrium is a function of the size of the population for a given mutation rate.
- There is a wide variance in expected outcomes for population genetic measurements for any given population history.

Continues on next page

- The expected time to most recent common ancestor TMRCA varies depending on population size and whether the population is expanding, contracting or remaining constant.
- Expanding populations have old TMRCA and lots of long branches
- Contracting populations have young TMRCA and few long branches.
- Constant sized populations have relatively deep TMRCA and lots of short branches
- The variance in expected TMRCA is equal to the square of the mean.
- This means that the distributions in expected TMRCA outcomes often overlap considerably for populations with extremely different histories
- Thus in order to distinguish between various population history scenarios we need to sample enough loci to estimate the observed shape of the distribution.
- Increasing sample size in terms of numbers of individuals increases the resolution for the inference of the TMRCA for individual genetic locus because you are more likely to capture the population TMRCA in the sample
- Increasing the sample size in terms of number of genetic lineages (loci) gives a better estimate in the variance in TMRCA between genetic lineages, and a better estimate of the overall shape of the distribution of TMRCA
- Demographic history can be inferred from even a single individual with complete genome data by looking at the genomic distribution of TMRCA between maternal and paternal chromosomes
- With a greater number of individuals fewer loci can be used, though multi locus data is always necessary.

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%)

Model Answer (Marker – james Rosindell (1st), Samraat Pawar (2nd)):

(i) ...

a. This is neutral because the individuals to die and reproduce are chosen in a way that does not depend on their species identity.

b. This is not neutral because the probability of reproduction for an individual would change depending on its species identity - a different species identity would likely mean a different number of conspecific individuals nearby.

c. This is neutral because although not all individuals have the same chance of death, the probability of death has nothing to do with species identity and is only dependent on spatial location.

Continues on next page

d. This is not neutral because some of the species react to the two different habitat types, but not all. So there must exist scenarios where an individual's prospects of death would be affected by switching its species identity from one that reacts to habitat type to one that does not.

(ii) ...

a. Dynamic equilibrium means some aspects of model (e.g. species richness) are stable whilst the species themselves are in a continuous sate of change. The burn in period is needed to ensure that dynamic equilibrium is reached before sampling data from the model (assuming it is the equilibrium behaviour we are interested in).

b. These are large simulations and only 800Mb of RAM has been requested, probably the RAM went over limit on initialisation and so the simulations were cut off.

c. These are large simulations meaning they proabably take a long time to burn on. A wall time of 2 hours probably means not enough time to burn in and so no useful results returned.

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
}
```

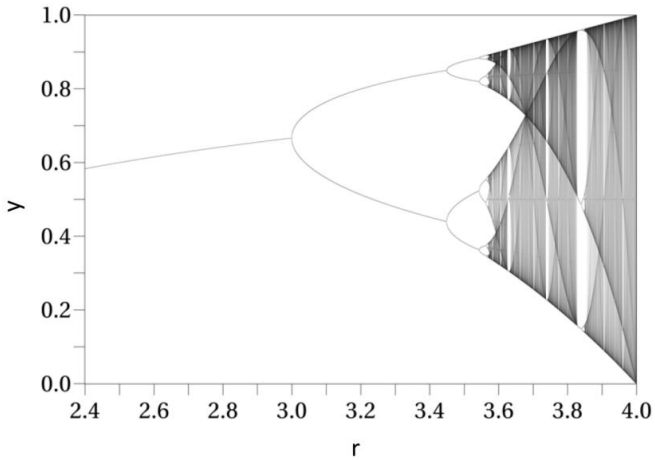
(i) Suppose $r = 2$, what value or values will get saved in the output. Explain your workings. (20 %)

(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 %)



Continues on next page

Model Answer (Marker – James Rosindell (1st), Samraat Pawar (2nd)):

(i) This should be possible without a calculator....

$$y_1 = 0.5$$

$$y_2 = 2 * 0.5 * (1 - 0.5) = 0.5$$

$$y_3 = 2 * 0.5 * (1 - 0.5) = 0.5$$

so all values are 0.5 no matter how many times we zip around the loops.

(ii) This should be possible without a calculator....

$$y_1 = 0.5$$

$$y_2 = 0.5 * 0.5 * (1 - 0.5) = 0.125$$

$$y_3 = 0.5 * 0.125 * (1 - 0.125) < 0.0625$$

$$y4 < 2^{-5}$$

so after running this n times when n is large, the remaining 100 values will be so close to 0 that really they might as well all be zeros.

(iii) This should be possible without a calculator....

$$y_1 = 0.5$$

$$y_2 = 8 * 0.5 * (1 - 0.5) = 2$$

$$y_3 = 8 * 2 * (1 - 2) = -16$$

$$y_4 = 8 * -16 * (1 - -16) = \text{some big negative number}$$

After this, it'll just get more and more negative so we can't say much about the values that'll get saved as it'll shoot quickly off to – infinity.

(iv) A pseudo random number generator is the outcome of a deterministic process that produces a set of apparently random numbers.

The random seed is a value used to initialize a pseudo random number generator. Using the same seed again will give you exactly the same sequence of pseudo random numbers.

(v) As r increases the sequence changes starts to flip between a pair of values and then between 4 different values. For larger r we get the onset of chaos meaning that a tiny change in the initial conditions (or in r) will yield a very large change in outcome. Any sensible large value of $r < 4$ will do for the random number generator.

End of paper