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: Wednesday, 09th Jan 2018, 14:00 – 17:00

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.

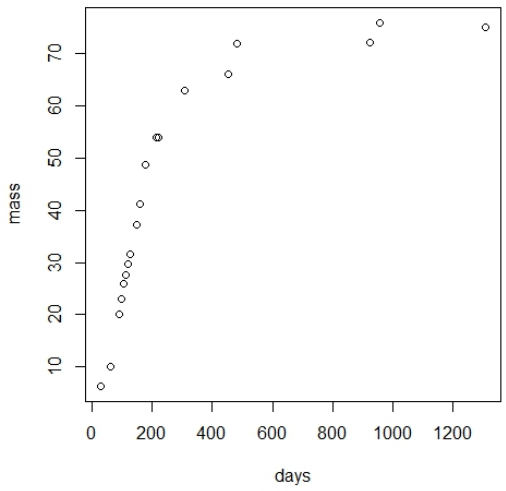
WE REALLY MEAN IT. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

Continues on next page

Section 1: 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 data on the growth (increase in biomass per unit area) of a population of Rhododendron after invasion into a new location in the northeast of England. Here is what the population's growth trajectory looks like:



Now answer the following:

(i) Name and describe at least one mechanistic and one purely phenomenological/statistical model you could fit to these data. Explain what biological mechanisms each model could/would capture, how you would determine which model among the two fits better, and the pros and cons of mechanistic vs. phenomenological modelling in this case. [70%]

(ii) Write out the appropriate R-/Python-/pseudo- code that would fit these models to the data. Explain what each command or code-block does with a single-line comment, as you would in the actual script. [30%]

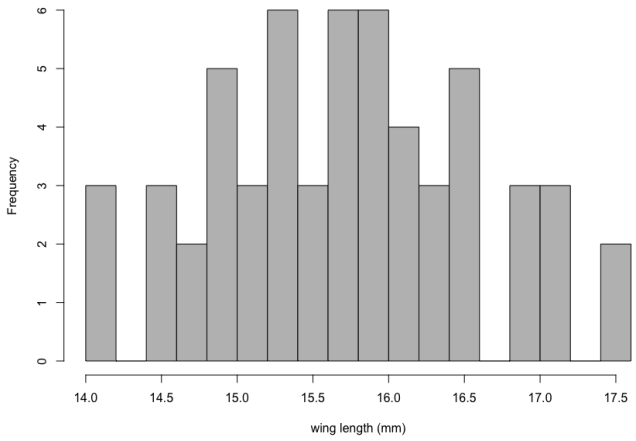
B. You are doing a research project on 57 bird skins that you found in the a natural history museum's collection. These birds are not catalogued so it is unclear what species they belong to. You believe that they are of the species *Parus lundyensis*. The species is clearly sexually dimorphic in plumage, allowing you to identify the sex of the skins. You want to find out whether these 57 bird skins are from the same species. You found a reference to that species in the "Handbook of the mysterious birds of the world" which states "*Parus lundyensis* shows a plumage dimorphism. Both sexes have a wing length of 15.5 – 17.0mm."

You conduct some exploratory data analysis, and then run the main test. Below is the R output from these analyses:

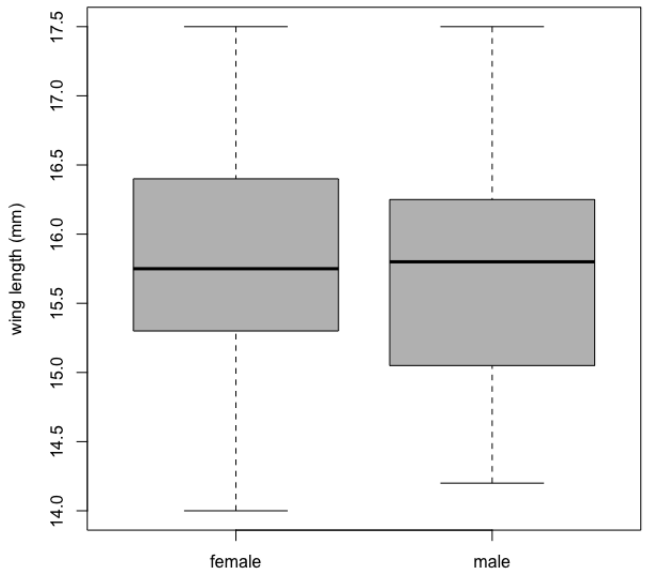
```
'data.frame':   57 obs. of  3 variables:
$ Catalogue_Nr   : int  1 2 3 4 5 6 7 8 9 10 ...
$ wing_length.mm.: num  16.5 15.2 16.1 16.6 14.8 16.6 15.9 15.8 15 15.9 ...
$ sex            : Factor w/ 2 levels "female","male": 2 1 2 1 2 1 1 2 1 1 ...

Catalogue_Nr wing_length.mm.    sex
1            1            16.5   male
2            2            15.2 female
3            3            16.1   male
4            4            16.6 female
```

Continues on next page

```
5            5            14.8   male
6            6            16.6 female
```




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.00   15.20   15.80   15.77   16.30   17.50
```




```
Welch Two Sample t-test

data:  data$wing_length.mm by data$sex
t = 0.074339, df = 55, p-value = 0.941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    -0.4422484  0.4763225
sample estimates:
mean in group female   mean in group male
            15.78000             15.76296
```

```
Welch Two Sample t-test
```

Continues on next page

```
data:  data$wing_length.mm by data$sex
t = 0.074339, df = 55, p-value = 0.941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
    -0.4422484  0.4763225
sample estimates:
mean in group female   mean in group male
            15.78000             15.76296


One Sample t-test

data:  data$wing_length.mm
t = -4.2154, df = 56, p-value = 9.175e-05
alternative hypothesis: true mean is not equal to 16.25
95 percent confidence interval:
    15.54474 15.99912
sample estimates:
mean of x
    15.77193
```

(i) Write a methods section detailing the aim, methods and statistical analysis you would do to achieve your aim. Fully justify your methods. [30%]

(ii) Write the R (or pseudo-) code that you need to produce the results starting with a blank R script (including the code to get the above results). Comment each line of code to explain what you do here and why. You can add code that would produce additional results that you think would be good to give. [35%]

(iii) Write a results section using the above output, including descriptive statistics, and effect sizes – quantify the outcome. If you come across issues, discuss them in writing. Write one concluding sentence at the end of this section that links back to the aim. [35%]

Continues on next page

Section 2: GIS, Genomics, C & Data structures

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

A. Using examples, describe the process by which satellite images are converted into data for analysis in a geographic information system [70%].

How does the spectral and spatial resolution of the image affect the accuracy of the data product [30%]?

B. In the lectures, we explored the use of binary node structures in C. Answer the following:

(i) Design a structure that could be used to represent a node with an arbitrary number of descendants. [30%]

(ii) Write a C function that uses recursion to traverse a tree constructed of such a node. Partial credit is given for correct pseudocode. [40%]

(iii) What are the safety issues associated with this node structure and traversal method? What are some ways of adjusting the structure or the function to mitigate these risks? [30%]

C. How and why does population structure affect expected genotype frequencies? Explain the relationship between the genotype frequencies and how these change with increasing population structure. How can genotype frequencies be used to test for population structure?

Continues on next page

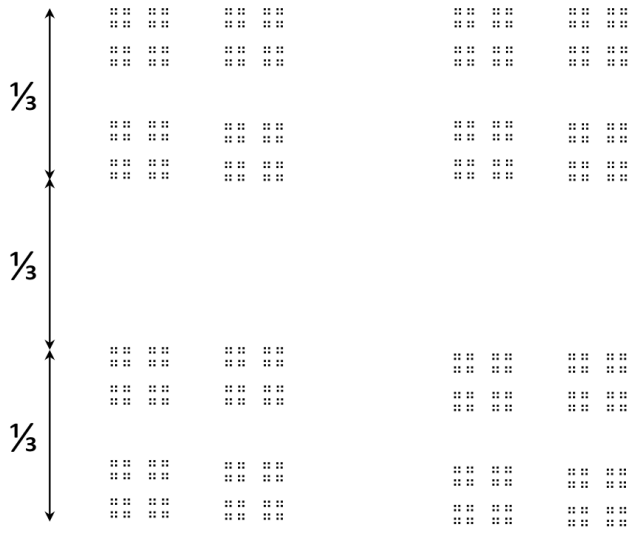
Section 3: Neutral theory HPC

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

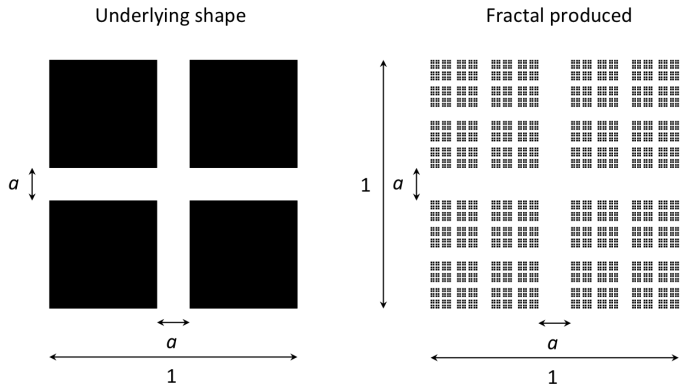
A. Answer the following questions. Please be brief in your answers.

(i) Give three reasons why fractals occur in the natural world giving an example for each. [30%]

(ii) Calculate the dimension of the following fractal, which is known as the Cantor Dust. Please show your workings. [20%]



(iii) The following diagram shows how to construct a range of different fractals depending on variable a where . Write down a formula for the dimension of the fractal in the general case, as a function of the variable a. The Cantor Dust shown in part b) corresponds to the case where $a = \frac{1}{3}$. Check that when you put $a = \frac{1}{3}$ into your formula you do get the same answer as you gave in b) [20%]



(iv) Write a short piece of pseudo code to draw this fractal as a function of a. In pseudo code you can say things like "draw a filled square of width w with the top left corner at `(x , y)`". *Hint*: define a function that calls itself four times and be careful to prevent the resulting loop from continuing infinitely. Your function should have five input parameters: a, w, x, y, threshold [30%]

B. Consider the following spatially explicit neutral model, also known as the voter model. In each time step, an individual is chosen at random to die and be replaced with the offspring of one of its eight immediate neighbours. With probability ν the new-born individual is of an entirely new species in the system (speciation) otherwise it is of the same species as it's parent. This question relates to performing simulations of such a model and observing its species richness.

(i) For the special case where $\nu = 0$, given an initial condition where there are many species in the

Continues on next page

system, describe what will happen to the diversity of the system as the simulation progresses and why. [20%]

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

(iii) You would need to repeat this stochastic simulation many times to get an idea of the overall behaviour of the system and you plan to do this using High Performance Computing (HPC). In the context of this goal describe the meaning of the following and their implications on performing your simulations:

a. The line of R code [20%]:

```
xvar <- as.numeric(Sys.getenv("PBS_ARRAY_INDEX"))
```

b. This line of shell script [20%]:

```
#PBS -l walltime=2:15:00
```

(iv) Suppose now that your simulation model is running on a fragmented landscape where not every position in space can be occupied. The simulation rules remain the same as before except that when a dead individual is replaced, there may not be eight immediate neighbours any more because some of the neighbouring spaces might not be suitable habitat. Consider the following fragmented landscape in which habitat is shown in black and non-habitat is shown in white (the boundary line is in black but does not count as habitat).

Imagine simulating an extremely low (practically zero) speciation rate with a voter model on this landscape, what species richness would you expect to see at dynamic equilibrium? Explain your answer. [20%]

End of paper