IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

LONG FORMAT EXAM

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

Exam Date: Tuesday, 30th March 2021, 10:00 – 13:00

Length of Exam: 3 HOURS

Instructions:

Please note that this exam has three Sections:

- Section 1 requires **ONE** of three questions to be answered
- Section 2 requires **ONE** of three questions to be answered
- Section 3 requires **ONE** of three questions to be answered

You must answer **three questions**, all of which have **equal weight**, so it is reasonable to spend about 1 hour on each question.

Read the instructions carefully at the head of each section. In addition:

- Please **follow the instructions for the remote assessment carefully**.
- Please check that you have **answered all the parts of each question** you choose - some questions cover more than one page.
- This is an open book exam: you may refer to teaching, revision materials and online resources but you **must not confer with any other individual during the examination**.
- You will have 30 minutes after the end of the exam to prepare and upload answer files. Make sure you upload a file for each question you answer to the correct question dropbox.

Continues on next page

Section A

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

Question A1 Bayesian Statistics

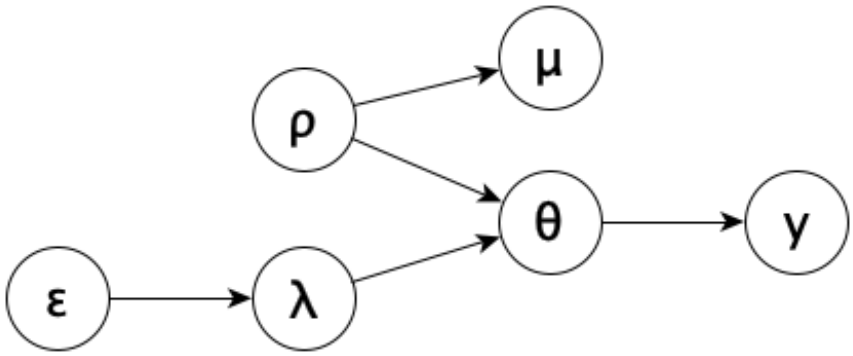
You wish to model the density of two species of bears in North America, namely the American black bear (*Ursus americanus*, labelled A) and the brown bear (*Ursus arctos*, labelled B). Assume that your random variable θ corresponds to the frequency of detecting species A, so that $\theta = [0, 1]$.

You gather results from a field survey conducted in one location in North America. The collected data y shows that out of 43 detected individuals, 30 were brown bears (species B) and 13 were American black bears (species A). Based on prior knowledge, you expect that, both on average and as the most probable value, the frequency of species A is 20% with a low-to-moderate confidence in this assertion.

(a) [50%] After choosing a suitable conjugate prior - likelihood Bayesian model, and after choosing appropriate values for the prior distribution based on the information provided, calculate the posterior mean of θ. Be as formal as possible, justify in words the choice of the parameters of the prior distribution, and provide brief but clear explanations of your calculations. Finally, provide the posterior mean of θ using a non-informative prior without extra calculations.

(b) [25%] Assume that you wish to model θ with a more complex likelihood function which can not be defined and evaluated. Nevertheless, you have a simulator function which is able to generate random draws of the number of detected bears of species A, y, out of the total number of detections, given a certain value of θ. You wish to use Approximate Bayesian Computation to derive the posterior ditribution of θ, calculate its posterior mean and 95% confidence intervals. Provide a well commented functioning code in a suitable programming language to perform this task.

(c) [15%] You now model the relationship between θ and several random variables that affect the observed data y with the following Bayesian network.



Using the chain rule of Bayesian networks, express the joint probability of $P(y, \theta, \lambda, \epsilon, \rho, \mu)$ as a product of conditional probabilities.

(d) [10%] Discuss how Bayesian networks are useful to model processes in ecology and evolution, using examples from the literature.

Continues on next page

Question A2 Generalised Linear Models

The adaptive significance of temperature-dependent sex determination (TSD) has attracted a great deal of research, but the underlying mechanisms by which temperature determines the sex of a developing embryo remain poorly understood. Here, we manipulated the level of a thyroid hormone (TH), triiodothyronine (T3), during embryonic development (by adding excess T3 to the eggs of the red-eared slider turtle *Trachemys scripta*, a reptile with TSD), to test whether TSD was affected by sex steroid hormones, such as T3. Our study has implications for the conservation of TSD reptiles in the context of global change because environmental contaminants may disrupt the activity of THs, and thereby affect offspring sex in TSD reptiles.

The dataset had the following structure:

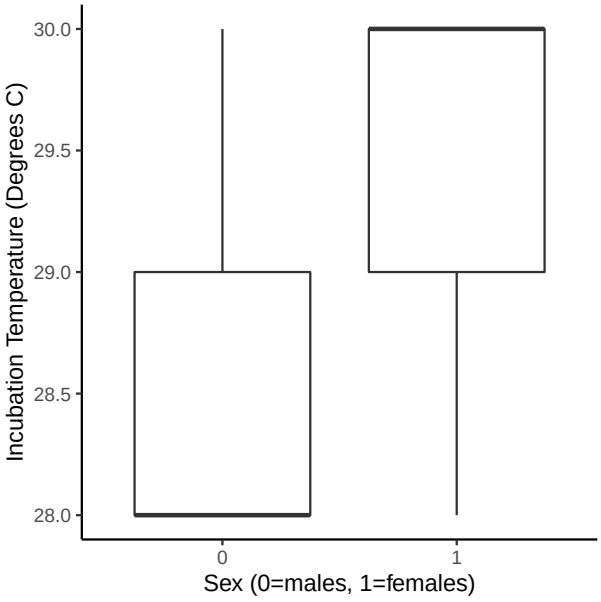
- IncuT = incubation temperature in degrees Celsius
- Treatment = factor of 2 levels (control and T3). T3 is a treatment where excessive thyroid hormone was applied, with control representing the baseline.
- Sex = sex of hatching individuals, males were encoded with 0 and females with 1

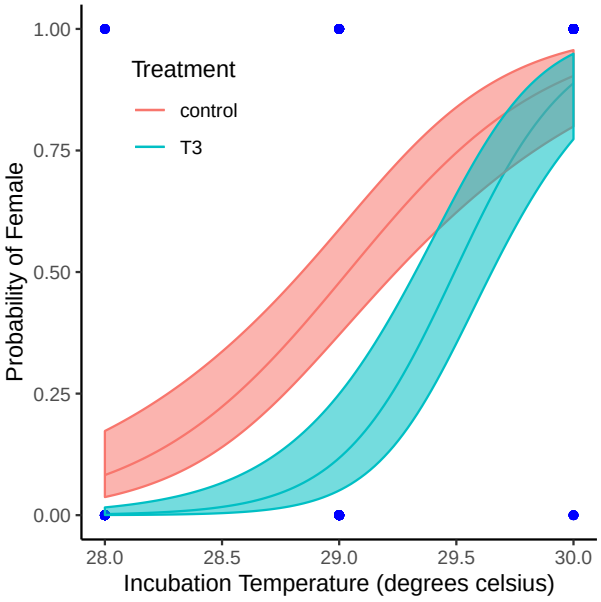
An analysis has been conducted to investigate the research question and the following results and R outputs have been obtained.

```
M1<- glm(Sex~IncuT*Treatment, data = data, family = "binomial")
summary(M1)
##
## Call:
## glm(formula = Sex ~ IncuT * Treatment, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.16221  -0.50002  -0.06667   0.48623   2.23398
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -67.4434    10.7454  -6.277 3.46e-10 ***
## IncuT               2.3226     0.3705   6.268 3.65e-10 ***
## TreatmentT3       -53.2357    21.3439  -2.494   0.0126 *
## IncuT:TreatmentT3   1.7692     0.7259   2.437   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 375.10  on 274  degrees of freedom
## Residual deviance: 191.19  on 271  degrees of freedom
## AIC: 199.19
##
## Number of Fisher Scoring iterations: 6

anova(M1, test = "Chisq")
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Sex
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             274     375.10
## IncuT           1  164.182       273     210.91 < 2.2e-16 ***
## Treatment       1   13.234       272     197.68  0.000275 ***
## IncuT:Treatment 1    6.488       271     191.19  0.010863 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Continues on next page





Using all the information above write:

(a) [30%] A statistical methods section, justifying the model structure and model family

(b) [70%] A results section summarising the outputs, including but not limited to describing the overall relationship, indicating the temperature at which being female is more likely, and estimating the goodness-of-fit of the model

Continues on next page

Question A3 Maximum Likelihood

Answer all three parts of the following question.

(a) [30%] Let $Y|\lambda \sim$ Exponential(λ), with conditional pdf $f_{Y|\lambda}(y|\lambda) = \lambda e^{-\lambda y}$. Here λ is also a random variable following another distribution, $\lambda \sim$ Uniform$(1, 2)$. Find the pdf of Y. Note that the pdf of Y does not depend on λ.

(b) [30%] Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from $N\left(0,\ \sigma^2\right)$. The parameter of interest is σ^2. Find the MLE for σ^2. You are given the pdf for X_i:

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} x_i^2}$$

(c) [10%] Using the results from part (b) or otherwise, find the MLE for σ.

(d) [30%] Discuss, in as much detail as possible, the inference of confidence intervals based on approximate normality (i.e. Wald confidence intervals). You may use graphs or equations to support your answers. Use θ to denote the parameter of interest and you may assume θ is univariate.

Continues on next page

Section B

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

Question B1 Maths 1

The Gompertz growth curve is sometimes used to study the growth of populations. Its properties are quite similar to the properties of the logistic growth curve. The Gompertz growth curve is given by

$$N(t) = K \exp[-ae^{-bt}]$$

for $t \geq 0$, where K and b are positive constants.

(a) [10%] Show that

$$a = \ln\left(\frac{K}{N_0}\right),$$

where $N_0 = N(0)$.

(b) [10%] Show that for $t \geq 0$, $N(t) < K$ if $N_0 < K$. (*Hint*: determine whether a and e^{-bt} are positive or negative in this case)

(c) [30%] Show that

$$\frac{dN}{dt} = bN(\ln K - \ln N)$$

and

$$\frac{d^2N}{dt^2} = b\frac{dN}{dt}[\ln K - \ln N - 1]$$

(d) [10%] Use your results in (b) and (c) to show that $N(t)$ is monotonically increasing if $N_0 < K$.

(e) [20%] Use your results in (b) and (d) to show that, if $N_0 < K$, the population size $N(t)$ tends to the value K if you wait long enough. What is the biological interpretation of K?

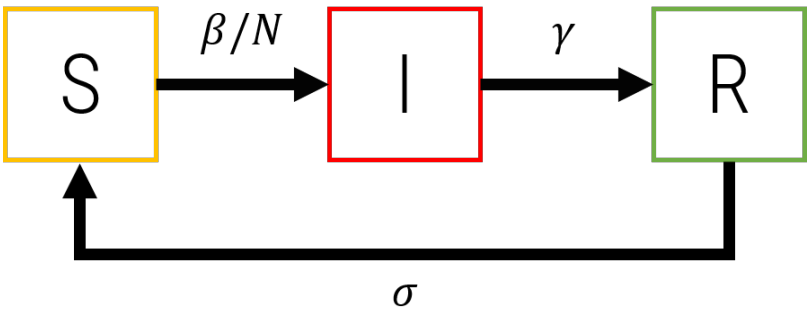
(f) [20%] Assume that $N_0 < K/e$ (where e is Euler's number). Show that $N(t)$ has an inflection point at $t_1 = \frac{1}{b}\ln a$.

Continues on next page

Question B2 Maths 2

The SIRS model of epidemiology is a set of coupled non-linear differential equations that describes the course of an infectious disease epidemic in a population of N individuals, with $S(t)$ susceptible, $I(t)$ infected, and $R(t)$ recovered (and immune) individuals. There are 3 events which can happen:

1) susceptible individuals come in contact with infected individuals causing new infections at rate β/N per pair-wise contact,
2) infectious individuals spontaneously recover at rate γ per individual and,
3) recovered individuals spontaneously loose immunity and become susceptible again at rate σ per individual.



This gives the following set of differential equations.

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta}{N}SI + \sigma R \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta}{N}SI - \gamma I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \sigma R \end{aligned}$$

(a) [15%] Using the fact that the population size is constant $N = S + I + R$, show that the 3 differential equations above can be reduced to the two equations below

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta}{N}SI + \sigma(N - S - I) \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta}{N}SI - \gamma I. \end{aligned}$$

Non-dimensionalise these equations by making the substitutions $x = S/N$, $y = I/N$ and $\tau = \gamma t$ to give

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}\tau} &= -R_0 xy + \eta(1 - x - y) \\ \frac{\mathrm{d}y}{\mathrm{d}\tau} &= (R_0 x - 1)y \end{aligned}$$

where $R_0 = \beta/\gamma$ and $\eta = \sigma/\gamma$.

(b) [15%] Show that the $\dot{x} = 0$ and $\dot{y} = 0$ nullclines of these equations are

$$\begin{aligned} y &= \frac{\eta(1 - x)}{R_0 x + \eta} \qquad (\dot{x} = 0) \\ x &= \frac{1}{R_0} \qquad (\dot{y} = 0) \end{aligned}$$

Continues on next page

excluding the $y = 0$ nullcline which corresponds to zero infections. Using this result or otherwise show the fixed points of the dynamics ($\dot{x} = \dot{y} = 0$) are

$$\begin{aligned} x^* &= \frac{1}{R_0} \\ y^* &= \frac{\eta(1 - \frac{1}{R_0})}{R_0 x + \eta}. \end{aligned}$$

Comment on the nature of the fixed points for $R_0 > 1$ and $R_0 \leq 1$ and the outcome of the epidemic in each case, should the fixed points be obtained.

(c) [20%] Assume that $R_0 > 1$. Sketch the nullclines on an $x - y$ axis. Using the nullclines and fixed points calculate the direction of the change on the nullcline ($\dot{x} \lessgtr 0$ on $\dot{y} = 0$ nullcline and $\dot{y} \lessgtr 0$ on $\dot{x} = 0$ nullcline) and indicate using arrows on your plot. From these flows on the nullclines we cannot say whether the fixed point is stable or unstable yet, but what type of dynamics do they imply for $R_0 > 1$?

(d) [20%] A linear stability analysis of the dynamics about the fixed point results in the following Jacobian matrix $\boldsymbol{J}$

$$\boldsymbol{J} = \begin{pmatrix} -\eta\frac{R_0+\eta}{1+\eta} & -(1+\eta) \\ \eta\frac{R_0-1}{1+\eta} & 0 \end{pmatrix}$$

For the case that $R_0 > 1$, make the assumption that the rate of loosing immunity σ is much smaller than the rate of recovery γ, to show that the Jacobian matrix is given approximately by:

$$\boldsymbol{J} \approx \begin{pmatrix} -\eta R_0 & -1 \\ \eta(R_0 - 1) & 0 \end{pmatrix}$$

and that its eigenvalues are

$$\lambda = \frac{-\eta R_0}{2} \pm \frac{1}{2}\sqrt{\eta^2 R_0^2 - 4\eta(R_0 - 1)}$$

Using this result determine whether the fixed point is stable?

(e) [15%] Using the eigenvalues show that the condition for the eigenvalues to be complex ($\mathrm{Im}(\lambda) \neq 0$) is

$$\eta R_0^2 < 4(R_0 - 1).$$

Hence, when this condition is satisfied what is the classification of this fixed point.

Assuming that this condition is satisfied with $R_0 > 1$, using this information sketch on the plot of nullclines the trajectory of the epidemic from the beginning to the end (Hint at the beginning of the epidemic $y \approx 0$, $y > 0$ and $x \approx 1$, but $x < 1$).

(f) [15%] If $\mathrm{Im}(\lambda) \neq 0$, by making the assumption that $\eta^2 R_0^2 \ll 4\eta(R_0 - 1)$, show that the angular frequency of the oscillations in radians/day is

$$\omega \approx \gamma\sqrt{\eta(R_0 - 1)}.$$

(Hint: if $\mathrm{Im}(\lambda) \neq 0$ then the dimensionless angular frequency is $\tilde{\omega} = \frac{\mathrm{d}\phi}{\mathrm{d}\tau} = \mathrm{Im}(\lambda)$, where ϕ is the phase angle oscillatory motion of x and y.)

For Sars-Cov2, we know that without any restrictions $R_0 \approx 2.5$ and $1/\gamma \approx$ 7days and that immunity probably lasts a few months, so for each of the following sets of parameters show that the condition for $\mathrm{Im}(\lambda) \neq 0$

Continues on next page

is satisfied and calculate the angular frequency of the oscillations ω in radians/day and the period of the oscillations $T = 2\pi/\omega$ in days (to the nearest day):

i) $R_0 = 2.5$, $\gamma = 1/7$ days^{-1} and $\sigma = 1/100$ days^{-1}

ii) $R_0 = 1.1$, $\gamma = 1/7$ days^{-1} and $\sigma = 1/100$ days^{-1}.

Continues on next page

Question B3 Evolutionary Modelling

A deterministic equation describing how the frequency x of a mutant changes when it has a selection coefficient s compared to the wild type is:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = sx(1-x)$$

(a) [15%] Comment on

i) the values of s for these continuous time dynamics to be a good representation of an evolutionary process with discrete generations?

ii) how this ODE is related to the logistic growth equation of a population with finite carrying capacity K.

iii) what assumption is made about the population size N, when using this ODE?

(b) [25%] Solve this ODE for the initial condition $x(t=0) = x_0$, using separation of variables, followed by partial fractions.

(c) [15%] If $s = -s_d$ $(s_d > 0)$ and $x_0 \ll 1/2$ explain why the following ODE is a good description of the dynamics

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -s_d x,$$

find the solution to this equation with initial condition x_0, and sketch the solution. Comment on long term fate (as $t \to \infty$) of the mutant given this solution.

(d) [10%] We can calculate an estimate of the time to extinction t^* by calculating the time to reach $x(t^*) = 1/N$, where N is the population size. Using this solution calculate t^*, given the initial frequency x_0, by setting the final frequency to $x(t^*) = 1/N$ in your solution. Comment on

i) why in the above calculation we cannot set the final frequency to $x(t^*) = 0$, and

ii) what aspect of this model would need to be rectified to properly calculate the time to extinction.

(e) [20%] In practice, mutations from wild type to mutant mean that the mutant does not go extinct. The following differential equation describes this for *unidirectional* mutation from wild type to mutant

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -s_d x + \mu(1-x)$$

Explain why the second term on RHS corresponding to mutation has the form that it does and solve this equation using the integrating factor method.

(f) [5%] Using this solution or otherwise show that the long-term or equilibrium ($\dot{x} = 0$) frequency x^* is given by

$$x^* = \frac{\mu}{\mu + s_d}$$

(g) [10%] Imagine a stretch of DNA of length $\ell = 100$kb. Making the *infinite sites/alleles* assumption, and assuming that all mutations that arise on the wild type haplotype/sequence are deleterious with the same selection coefficient $s_d = 0.0001$ and that the nucleotide mutation rate of $\mu = 10^{-9}$, calculate, using this result, the equilibrium frequency of all mutants in the population that are not wild type. (Hint: assume that each new sequence on wild type that arises segregates independently and that no mutants arise on the background of any mutant sequence).

Continues on next page

Section C

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

Question C1 Ecological Modelling

The Iron Hypothesis states that growth of phytoplankton in the oceans is limited by the availability of iron. By introducing extra iron to the ocean waters, the growth of plankton and their total abundance can increase and, consequently, the uptake of carbon dioxide would also increase.

This hypothesis was tested by a team of researchers in the IronEx I experiment. They found that the addition of iron to the ocean water caused an initial doubling of the amount of plankton, and the rate of growth quadrupled. However, after one day, the phytoplankton activity levelled off. This coincided with an increase in the amount of zooplankton, which feed on phytoplankton. The effects of the addition of iron were less than what was predicted on the basis of the response of phytoplankton alone (i.e. in the absence of zooplankton).

(a) [20%] What is a suitable model to describe this interaction?

(b) [20%] Draw a bifurcation diagram that shows the equilibria of this model and their stability

(c) [50%] Using this bifurcation diagram diagram explain the result from the IronEx I experiment.

Continues on next page

Question C2 GIS and spatial modelling

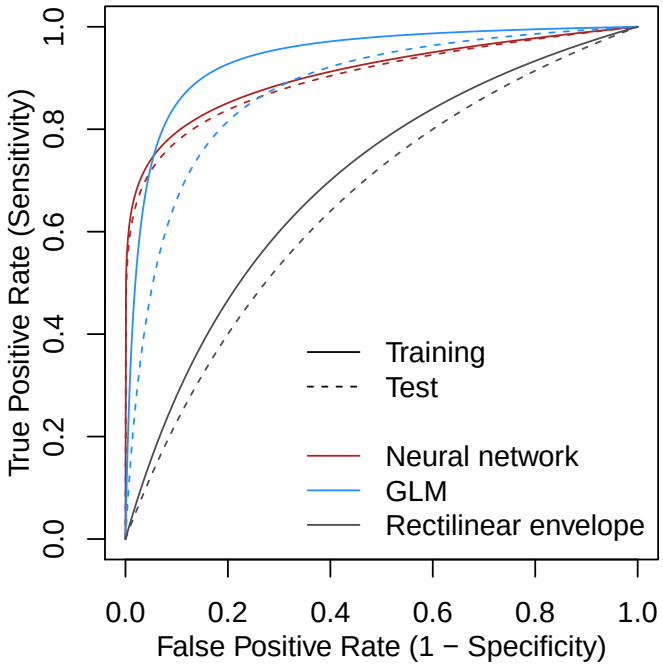
You are working with colleagues to create a species distribution model for the Fantastic Monkey, which is found in the warmer and drier lowlands of the UK. You have the following data:

- The Ordnance Survey Terrain 50 digital elevation model: a raster dataset in the OSGB36 projection with a 50 metre resolution.
- The 19 bioclimatic variables calculated from the Worldclim historical (1970 – 2000) climate data. This is a global raster at 30 arc-second resolution in the WGS84 projection.
- A CSV file of 1,150 sightings from the UK Fantastic Monkey Survey Group. These contain GPS locations as WGS84 latitude and longitude for confirmed sightings across the UK.

(a) [30%] Describe the data handling and GIS steps you would need to take to prepare this data for use in the species distribution model.

(b) [30%] The Fantastic Monkey Survey Group is requesting that you evaluate three model types: a rectilinear envelope, a neural network and a GLM. Provide feedback to the group on the advantages and disadvantages of those approaches.

(c) [40%] Fitting the three models above to a training dataset and predicting for a test partition gives the following receiver operating characteristic (ROC) curves. Explain what these curves mean and what the group should understand about the relative performance of the three models.



Continues on next page

Question C3 Genomics

This question is divided into four points, each one contributing equally to the whole grade.

Let us assume two subpopulations, each one in Hardy-Weinberg equilibrium. At one locus, the frequency of allele A in each subpopulations is f_{A1} and f_{A2}, respectively. The proportion of heterozygous individuals HS is equal to:

$$HS = (2f_{A1}(1 - f_{A1}) + 2f_{A2}(1 - f_{A2}))/2$$

On the other hand, the expected proportion of heterozygous individuals in a population with frequency $f_A = (f_{A1} + f_{A2})/2$, called HT, is equal to $HT = HS + (D^2)/2$ with $D = |f_{A1} - f_{A1}|$. From these quantities, we can calculate an estimate of the population subdivision FST equal to $(HT - HS)/HT$.

(a) [25%] Write a function in R (or any other suitable programming language) that returns the value of HT, HS, FST when giving the frequency of allele A in each subpopulation as input.

(b) [25%] Calculate FST when the frequency of allele A in subpopulation 1 and 2 are:

- $f_{A1} = 0.2$ and $f_{A2} = 0.1$,
- $f_{A1} = 0.3$ and $f_{A2} = 0.3$, and
- $f_{A1} = 0$ and $f_{A2} = 1$.

(c) [25%] Discuss the results obtained in points (b) in regards to the whole population being structured or not. What factors may influence levels of population structure at the genomic level?

(d) [25%] Discuss the use of FST to test for adaptation and illustrate the rationale using examples from the literature.

End of paper