IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

EXAM 2

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

Exam Date: Wednesday, 27th March 2017, 10:00 – 13:00

Length of Exam: 3 HOURS

Instructions: All sections are weighted equally. It is a three-hour exam, and there are 5 sections, so it is a reasonable guideline to spend about 35 minutes on each section. Most sections allow you to choose between questions, answering ONE. Please read the instructions at the head of each section carefully.

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

WE REALLY MEAN IT. PLEASE PUT ANSWERS TO EACH SECTION IN A SEPARATE EXAM BOOK. 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: Maths

A. Solve **one of the following two problems** [30%]. Please indicate clearly in your answer book which question you are answering (i.e., A(i) or A(ii)).

(i) Solve the following integral:

$$\int \frac{dx}{x^2 - 4x + 3}$$

(ii) Obtain the order three expansion of the function $f(x) = \frac{\ln x}{x}$ at $x_0 = 1$.

Model Answer (Marker – Bhavin Khatri (1st), James Rosindell (2nd)):

(i) When we find rational functions we should simplify them first with the methods we learned. In this case, the degree of the denominator is larger than that of the numerator. Decomposing it in simple fractions yields

$$\frac{1}{x^2 - 4x + 3} = \frac{1}{(x-3)(x-1)} = \frac{A}{(x+3)} + \frac{B}{(x-1)}$$

Leading to the following equations for the coefficients:

$$A = \left(\frac{1}{x-1}\right)_{x=3} = \frac{1}{2}$$
$$B = \left(\frac{1}{x-3}\right)_{x=1} = \frac{1}{2}$$

We have now easy integrals to solve:

$$I = \int \frac{x^3 - x}{(x-3)(x-1)} dx = \frac{1}{2}\int \frac{1}{x-3} dx - \frac{1}{2}\int \frac{1}{x-1} dx$$

$$= \frac{1}{2}\log|x-3| - \frac{1}{2}\log|x-1| + C$$

And we can use logarithmic relationships to obtain a clean result:

$$I = \log\sqrt{\left|\frac{x-3}{x-1}\right|} + C.$$

(ii) We start computing derivatives:

$$\begin{aligned} f'(x) &= \frac{1 - \log x}{x^2} \\ f''(x) &= \frac{2\log x - 3}{x^3} \\ f'''(x) &= \frac{11 - 6\log x}{x^4} \end{aligned}$$

Now we evaluate the function and the derivatives at the point requested: $f(1) = 0$, $f'(1) = 1$, $f''(1) = -3$ and $f'''(1) = 11$. And the polynomial will be:

$$T_2(1) = (x-1) - \frac{3}{2}(x-1)^2 + \frac{11}{6}(x-1)^3.$$

Continues on next page

B. Solve **one of the following two problems** [70%]. Please indicate clearly in your answer book which question you are answering (i.e., B(i) or B(ii)).

(i) Solve the differential equation:

$$y' - y/2 = 2\sin(3t)$$

and explain if the general solution is finite when $t \to \infty$.

(ii) Diagonalize the following matrix A, and find the matrix P such that $D = P^{-1}AP$, being D the diagonal matrix. Once you obtain D and P, demonstrate that $A = PDP^{-1}$. Finally, find the value of A^5.

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}$$

Model Answer (Marker – Bhavin Khatri (1st), James Rosindell (2nd)):

(i) The integrating factor will be

$$\mu(t) = e^{-\int \frac{1}{2} dt} = e^{-\frac{t}{2}}$$

that multiplies both sides of the equation, which yields

$$e^{-\frac{t}{2}}y' - e^{-\frac{t}{2}}y/2 = e^{-\frac{t}{2}}2\sin(3t).$$

As usual, we double check that the left hand side corresponds with the differential we are looking for

$$d(e^{-\frac{t}{2}}y) = e^{-\frac{t}{2}}y' - e^{-\frac{t}{2}}y/2,$$

thus

$$d(e^{-\frac{t}{2}}y) = e^{-\frac{t}{2}}2\sin(3t)$$

that we can integrate. The right hand side should be integrated by parts twice, [APG: This could be provided as a clue for the students] which yields

$$y(t) = -\frac{24}{37}\cos(3t) - \frac{4}{37}\sin(3t) + ce^{\frac{t}{2}}.$$

An interesting observation of this solution is that its behaviour when $t \to \infty$ depends on the sign of c [APG. that the behaviour depends on the sign of the constant could be provided as a clue]. If $c = 0$ the solution is finite but, as soon as it is different from zero, it will tend to ∞ if $c > 0$ and to $-\infty$ if $c < 0$.

(ii) We start finding the roots of the characteristic polynomial $\det(A - \lambda I) = 0$:

$$\begin{vmatrix} 2-\lambda & 2 \\ 1 & 3-\lambda \end{vmatrix} = \lambda^2 - 5\lambda + 4 = 0.$$

We get the roots $\lambda_1 = 1$ and $\lambda_2 = 4$, which are the eigenvalues of A. To find the first eigenvector we subtract λ_1 from the diagonal of A:

Continues on next page

$$M = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

and we solve the homogeneous system of equations $Mx = 0$, which obviously leads to $-x = 2y$ and thus $v_1 = (2a,\ -a)$. Making $a = 1$, it yields $v_1 = (2,\ -1)$. Proceeding similarly for $\lambda_2 = 4$ you should obtain the vector $v_2 = (a,\ a)$, and we pick up as particular solution $v_2 = (1,\ 1)$. Therefore, the matrix of the eigenvectors will be the matrix of change of basis

$$P = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}.$$

If we now compute the inverse of P using the formula $P^{-1} = \frac{1}{\det P}\text{adj}(A)$ we get

$$P^{-1} = \frac{1}{3}\begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} -$$

You can demonstrate that the diagonal matrix D, that we already know after computing the eigenvalues of A

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix},$$

can be obtained with the matrix P and its inverse with the formula $D = P^{-1}AP$. And the other way around, we can obtain A from D making $A = PDP^{-1}$. We finally compute $A^5 = PD^5P^{-1}$

$$A^5 = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1024 \end{pmatrix}\frac{1}{3}\begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 342 & 682 \\ 341 & 683 \end{pmatrix}.$$

Continues on next page

Section 2: Dynamical Models in Ecology

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

A. Alice studies competition between two species. The two species' diets overlap to a large extent. The animals she studies search for their food, mainly nuts and seeds, and once they have found food items, they spend a considerable time opening and consuming these. She wants to implement a mathematical model for the interaction between the two species, and makes the following logical arguments/assumptions:

- The more individuals there are, the fewer food items there will be.
- Animals of species 1 are on average twice the size of species 2.
- She assumes that the density of food items (F) depends linearly on the density of species 1 (N_1) and species 2 (N_2) as
$$F = F_0 - 2N_1 - N_2$$
- The intake of food items follows a Holling type II functional response, and because of the size differences, the handling times (h_1, h_2) and search rates for food items (α_1, α_2) differ between the two species. Given the size differences, the growth rate (r_1, r_2) and mortality rates (m_1, m_2) are also different.

With these considerations, she writes down the following differential equation model for the changes in the densities of species 1 and 2:

$$\frac{dN_1}{dt} = r_1 N_1 \left(\frac{\alpha_1 F}{1 + \alpha_1 h_1 F} - m_1 \right)$$

$$\frac{dN_2}{dt} = r_2 N_2 \left(\frac{\alpha_2 F}{1 + \alpha_2 h_2 F} - m_2 \right)$$

She wants to know what the outcome of the competition between these two species

(i) Draw the isoclines for this model in the phase plane. Based on this, predict the outcome of the competition [25%]

(ii) Alice needs to explain the results of her model to an ecologist. Explain her results in ecological terms [30%]

(iii) The ecologist asks if the outcome of competition would have been different if she would have used different parameters. When is the outcome of the competition different? [30%]

(iv) The ecologist asks if the result would have been qualitatively different if she would not have used a different functional response than the Holling type II. Would it? [15%]

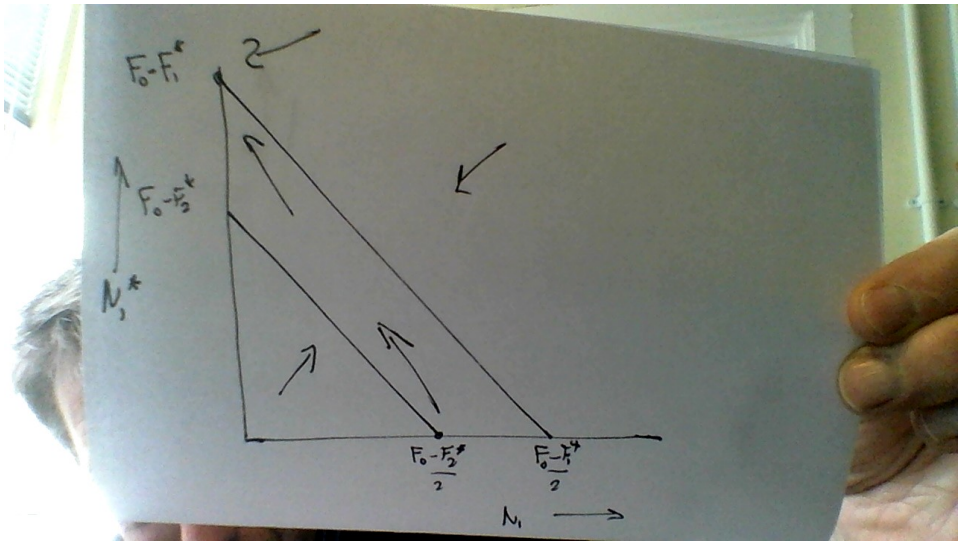
Model Answer (Markers – Bhavin Khatri (first), James Rosindell (second)):

(i) The key to answering the questions is that the isoclines are given by the two lines

$$N_2 = F_0 - F_i * -2N_1,$$

with $F_i* = m_i/((a_i(1 - h_i m_i)))$.

Continues on next page



These are straight lines in the phase plane, going through the points $(0, F_0 - F_i*)$ and $((F_0 - F_i*)/2, 0)$. The lines do not intersect (unless the F_i* happen to be the same, in which case they completely overlap) and therefore there is no equilibrium. One species excludes the other. From on invasion analysis in the points where one species is absent it follows that the winner of the competition is the one for which F_i* takes the lowest value. (This last part can also be given as answer in c. So if it is observed that this is always competitive exclusion that is sufficient for answering this question)

(ii) One species excludes the other. The simple reason is that they share a single resource, and therefore they are occupying the same niche.

(iii) No matter which parameters are chosen, one species will always exclude the other (partial marks if that is the only answer). If

$$\frac{m_1}{\alpha_1(1-h_1 m_1)} < \frac{m_2}{\alpha_2(1-h_2 m_2)},$$

species 1 is the winner of competition, otherwise species 2.

(iv) This is to make it a bit harder, and a bit more abstract. The answer is, no, that wouldn't matter. For a different functional response the winner of the competition could be different, but there will still be exclusion. Coexistence, or an outcome with alternative stable states is not a possibility. In ecological terms: they are still competing for one resource. In mathematical terms: let the functional responses be given by $g_1(F)$ and $g_2(F)$. The isoclines are given by $F_i* = g_i^{-1} m_i$, where g_i^{-1} is the inverse of g_i. No matter what is chosen for the functional responses, at the isoclines there will be a single value for F_i*. The isoclines will still be parallel straight lines that cannot intersect. In fact, you could even define a different relationship for the resource and still find that the isoclines might not be straight lines anymore, but that as long as there is a single resource that cannot intersect (But that is not needed for a correct answer. Seeing that the answer is no, and giving a credible argument suffices).

B. An ordinary differential equation (ODE) to describe the logistic growth of a population of fish, where at a constant rate h fish are hunted is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = vx\left(1 - \frac{x}{x_{\max}}\right) - hx$$

where x is the population size, v is the growth rate and $x_{\max}$ is the maximum population or carrying capacity in the absence of harvesting. Assume that $x \geq 0$ (i.e. that population cannot be negative) and $h \geq 0$.

(i) Put this equation in non-dimensional form,

Continues on next page

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = z(1-z) - \phi z = (1 - z - \phi),$$

using $\tau = vt$ and $z = x/x_{\max}$, where you should find $\phi = h/v$. Interpret what this constant ϕ means. [15%]

(ii) Determine the fixed points of this ODE for, (a) $\phi = 0$ (normal logistic growth), (b) $0 < \phi < 1$, & (c) $\phi = 0$. [15%]

(iii) Sketch the phase portrait of this ODE for (a), (b), & (c) , by plotting the RHS of above equation as function of z and indicating the stability of each fixed point by drawing appropriate arrows on the z-axis of the plot and using filled or open circles, respectively, for stable and unstable fixed points. [35%]

(iv) For each of (a), (b) and (c), answer whether there is always a stable population independent of the initial conditions (where $z(\tau = 0) > 0$). [15%]

(v) Using these results write an expression for the maximum stable population x^* as a function of the parameters $h, v, x_{\max}$. Describe qualitatively how the maximum stable population changes as h is increased. What is the critical value of $h^\dagger$, for which there is no longer a stable population. Interpret this biologically. [20%]

Model Answer (Markers – Bhavin Khatri (first), James Rosindell (second)):

Continues on next page

ANSWERS $\frac{dn}{dt} = vn\left(1-\frac{n}{n_{max}}\right) - hn$

a) $\tau = vt$ & $z = n/n_{max}$.

Substitute $\tau = vt$ on LHS

$$\frac{dn}{dt} = \frac{dz}{dt}\frac{dn}{dz} = v\frac{dn}{dz}$$

Substitute $z = n/n_{max}$ on LHS

$$v\frac{dn}{dz} = n_{max}v\frac{dz}{dz}$$

& RHS

$$vn\left(1-\frac{n}{n_{max}}\right) - hn = n_{max}vz(1-z) - n_{max}hz$$

∴ LHS = RHS

$$n_{max}v\frac{dz}{dz} = n_{max}(v z(1-z) - hz)$$

$$\therefore \frac{dz}{dz} = z(1-z) - \frac{h}{v}z$$

$$= z(1-z) - \phi z = z(1-\phi-z)$$

$\phi = h/v$ is the ratio of the hunting(reward) rate to the growth rate of the fish.

b) i) $\phi = 0$: $\frac{dz}{dt} = z(1-z)$

fixed points are : $z_1 = 0$ & $z_2 = 1$.

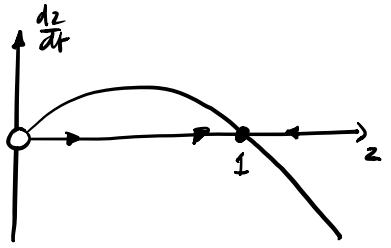
ii) $0 < \phi < 1$: $\frac{dz}{dt} = z(1-\phi-z)$

fixed points are : $z_1 = 0$ & $z_2 = 1-\phi$

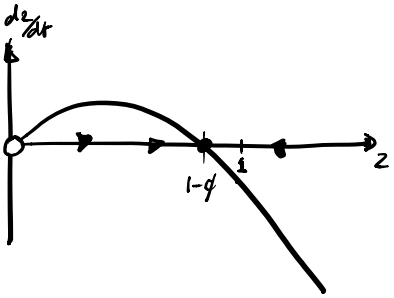
iii) $\phi = 1$: $\frac{dz}{dt} = -z^2$

Single fixed point at $z_1 = 0$

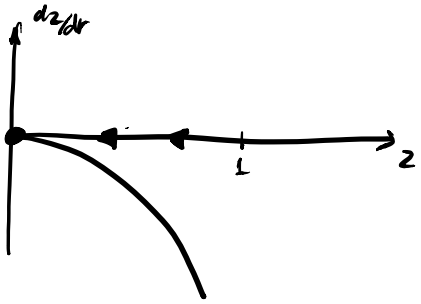
c) i) $\phi = 0$



ii) $0 < \phi < 1$



iii) $\phi = 1$



* As values of $x < 0$ or $z < 0$ are not possible effectively this is a stable fixed point. & not a half-stable fixed point ●

d) i) & ii) Yes, as for any value of z (or x), where $z(0) \neq 0$ have phase-portraits which lead to the fixed point (z_2) which is non-zero.

iii) No, as the phase portrait leads all solutions to fixed point at $z = 0$.

e) $z_2 = 1 - \phi$ $\Rightarrow x^* = x_{max} z_2$

$= x_{max}(1 - \phi)$

$= x_{max}\left(1 - \frac{h}{v}\right)$

* As h is increased the maximum stable population decreases linearly.
* The critical value of h is $h^* = v$, which means $x^* = 0$ & there is no longer a stable population.
* Biologically this corresponds to the hunting rate equalling the growth rate of the fish.

Section 3: Population Genetics & Evolutionary Ecology

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

A. You are in charge of a chemostat facility growing single-celled freshwater algae (*Chlorella* sp.) for an animal feed company. The algae grow in classic illuminated chemostats with a flow through of nutrient medium – the growth rate of the algae is limited solely by the concentration of nitrate in the inflow, every other nutrient, light and CO_2 is in excess. The system can be modeled by the following ODE system:

$$\frac{dN}{dt} = \frac{ckNS}{m+S} - DN$$
$$\frac{dS}{dt} = D(S_i - S) - \frac{kNS}{m+S}$$

(i) Explain each term in the model, taking care to define the parameters and give indicative units for each of them, including the variables. [25%]

(ii) Algae are harvested by filtering them out of the out-flow from the chemostats. What is the steady-state rate of production of algal biomass? What would you alter in the system in order to increase the rate of production? [25%]

(iii) You are thinking of increasing the temperature of the chemostats in order to speed up the production rate further. Sketch out the model could be extended to predict how the algae will respond to an increase in temperature. Use diagrams and/or equations where useful for your explanations [50%]

Model Answer (Marker – Tim Barraclough (1st), Austin Burt (2nd)):

(i) c is change in biomass per mole of nitrate 'handled' k is rate of processing of nitrate per g of alga N is biomass of alga in g/ml S is concentration of nitrate in chemostat moles/ml m is half saturation constant in moles/ml D is dilution rate as percentage

(ii) Solve equations to get steady-state

$$\hat{S} = \frac{Dm}{yck - D}$$
$$\hat{N} = c(S_i - \hat{S})$$

The production rate is N times flow rate in ml/hour => if they use D for this, would be production rate per ml of outflow per second, multiply by total volume of outflow for total production rate. Merit would solve equations but without insight of how to apply here.

(iii) Didn't cover temperature in my lectures, but have done in other parts of course. A suitable answer would be to make the growth rate temperature dependent, i.e. $k = exp(-E/kt)$ would be increase according to Boltzmann, or with drop again at higher temperatures would be e.g. Sharpe-Schoolfield equation. So assuming not already close to optimum temperature, would increase with decelerating exponential. But if there is mutational variation for temperature-dependence of growth, will get selection acting as well and if temperature is really totally fixed, might expect evolve so optimum matches the system temperature. They might sketch out an equation for this which would be something like (I am using a simplified version with Gaussian function for growth around optimum temperature):

$$k = e^{-(T-T_{opt})^2/\sigma}$$

(T = temperature, T_{opt}= optimum T for growth, σ = ecological tolerance around optimum)

Fitness,

Continues on next page

$W = ce^{-(T-T_{opt})^2/\sigma}\frac{S}{m+S} - D$

Change in T_{opt} per unit time

$$\frac{dT_{opt}}{dt} = \mu\frac{dW}{dT_{opt}}$$

where μ is mutation rate. Don't expect them to calculate all this.

They might also mention the possibility of assimilation/plastic response in addition to genetic adaptation. Merit = good verbal outline of theory, plus useful figures. Distinction = bring together detail on how to model evolution in such cases with the metabolic theory, evidence of synthesizing material from different parts of course and applying to example here.

B. Answer the following:

(i) Suppose you have written a program in `R` to simulate the change in allele frequency over time due to genetic drift. The simulator takes three input arguments: p_0, the initial allele frequency of a particular allele; N, the effective population size; and t, the number of generations to be simulated forward in time. The simulator returns the allele frequencies over successive generations. Explain, in as much detail as possible, how you can use this simulator to express the probability of fixation of an allele before $t = 30$ generations, as a function of p_0 and N. [40%]

(ii) (a) Suppose there is an infinite, random-mating population with 2 alleles, A and B. The frequency for the two alleles at generation t is p_t and q_t respectively, with $p_t + q_t = 1$. The three diploid genotypes AA, AB, and BB are at Hardy-Weinberg equilibrium. Furthermore, the AB heterozygotes produce A gametes in proportion d, and B gametes with proportion $1 - d$. Show that, in the next generation, the allele frequency for A is $p_{t+1} = p_t^2 + 2p_tq_td$. [30%]

(b) Show that, if $d > 0.5$, then the allele frequency of A is strictly increasing (i.e. $p_{t+1} > p_t$), as long as A remains polymorphic. [30%]

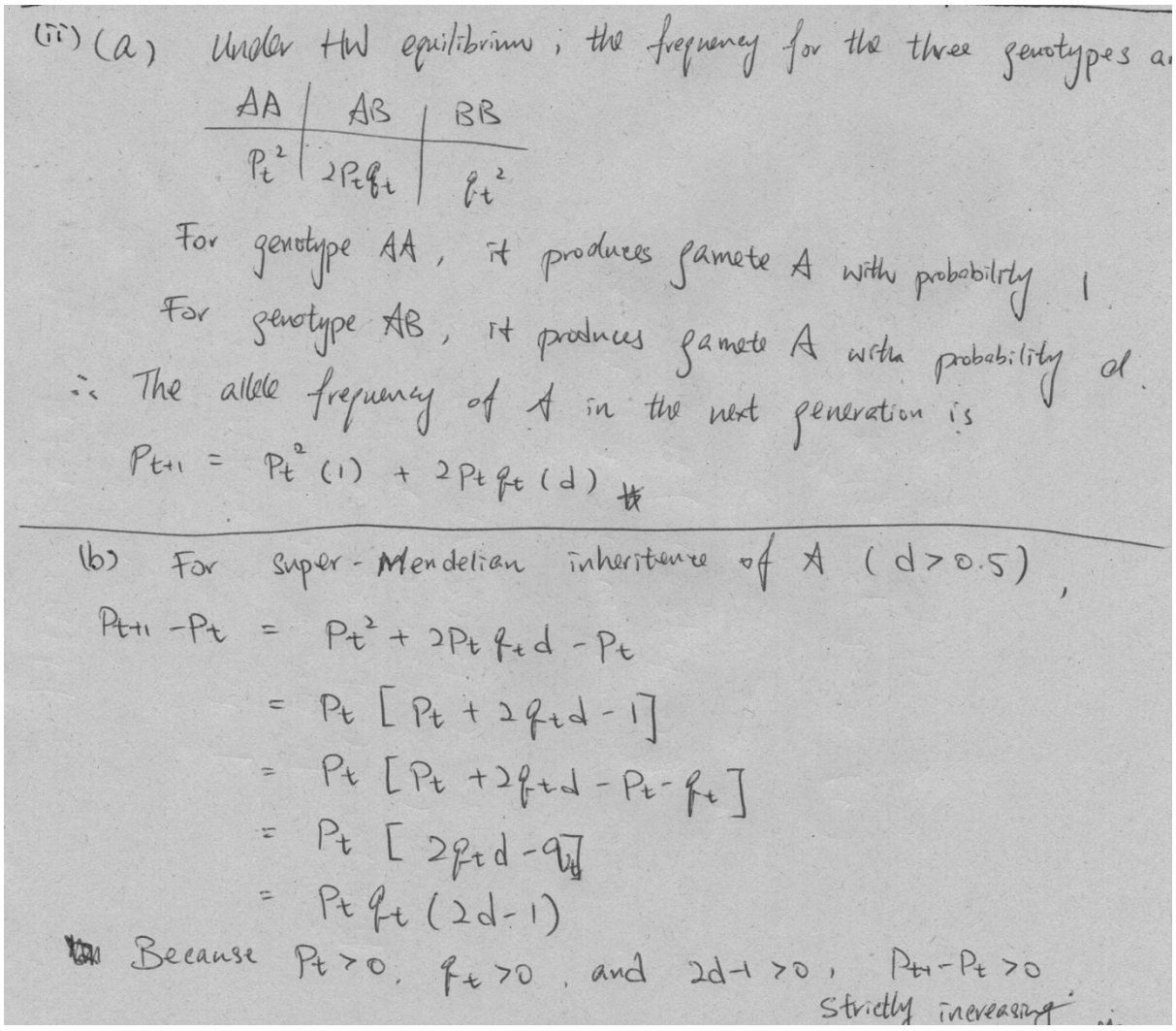
Model Answer (Marker – Austin Burt (1st), Tin-yu Hui (2nd)):

Continues on next page

(i)

(A similar exercise was discussed in class to demonstrate the use of Monte Carlo simulation. Students may use R/pseudo computer codes as part of explanation.)

- Set a range of values of p_0 (say, from 0.01 to 0.99, with 0.01 interval)
- Set a range of values of N_e (say, from 10 to 1000, with any reasonable interval)
- For each pair of p_0 and N_e combination, run the simulators for 10000 times (say) to obtain 10000 different allele frequencies after $t=30$ generations
- Count the proportion of having allele frequency of 1 (fixation) out of these 10000 simulations. And this would be your estimate for the probability of fixation.
- Repeat step 3 and 4 for other values of p_0 and N_e



(ii) (a) Under HW equilibrium, the frequency for the three genotypes a

AA	AB	BB
P_t^2	$2P_tq_t$	q_t^2

For genotype AA, it produces gamete A with probability 1.

For genotype AB, it produces gamete A with probability of .

∴ The allele frequency of A in the next generation is

$$P_{t+1} = P_t^2(1) + 2P_tq_t(d)$$

(b) For super-Mendelian inheritance of A ($d>0.5$),

$$\begin{aligned} P_{t+1}-P_t &= P_t^2 + 2P_tq_td - P_t \\ &= P_t[P_t + 2q_td - 1] \\ &= P_t[P_t + 2q_td - P_t - q_t] \\ &= P_t[2q_td - q_t] \\ &= P_tq_t(2d-1) \end{aligned}$$

Because $P_t>0$, $q_t>0$, and $2d-1>0$, $P_{t+1}-P_t>0$

Strictly increasing

Section 4: Maximum Likelihood & GLMs

Please select exactly **one question** and answer it. Calculator may be required in some questions.

A. (i) Let Y be a uniform random variable with lower bound a and upper bound b. The probability density function of Y is $f(Y) = \frac{1}{b-a}$.

(a) Calculate $E[Y]$ [20%]

(b) Calculate $E[Y^2]$. Then show that $Var[Y] = \frac{(b-a)^2}{12}$ [25%].

Hint: $b^3 - a^3 = (b-a)(a^2 + ab + b^2)$

(ii) Let $X_1, X_2, ..., X_n$ be i.i.d. Exponential random variables with rate λ. The probability density function of an Exponential random variable X is $f(X) = \lambda e^{\lambda x}$.

(a) Show that the MLE for λ is $\frac{n}{\sum_{i=1}^{n} x_i}$. [30%]

(b) Suppose we have five observations, 0.28, 0.66, 0.56, 0.43, and 1.07. Compute the MLE for λ and its 95% confidence interval based on approximate normality. [25%]

Model Answer (Markers – Tin-yu Hui (first), Austin Burt (second)):

Continues on next page

(ii) (a) $Y \sim \text{Uniform}(a,b)$, $f(y) = \frac{1}{b-a}$, $b > a$

Support of Y: (a,b)

$$E(Y) = \int_a^b y \frac{1}{b-a}\, dy = \frac{1}{b-a}\int_a^b y\, dy = \frac{1}{b-a}\left[\frac{y^2}{2}\right]_a^b$$

$$= \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{a+b}{2}$$

(b)
$$E(Y^2) = \int_a^b y^2\left(\frac{1}{b-a}\right)dy = \frac{1}{b-a}\int_a^b y^2\, dy = \frac{1}{b-a}\left[\frac{y^3}{3}\right]_a^b$$

$$= \frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right) = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2$$

$$= \frac{4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12}$$

$$= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12}$$

$$= \frac{b^2-2ab+a^2}{12}$$

$$= \frac{(b-a)^2}{12}$$

(ii)

(a) $L(\lambda) = f(x_1, x_2, \cdots, x_n)$

$= \prod_{i=1}^{n} f(x_i)$ (independent)

$= \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$ (identically distributed)

$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$

$l(\lambda) = \log[L(\lambda)] = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$

$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$

Find a $\hat{\lambda}$ s.t. $l'(\hat{\lambda}) = 0$

i.e. $\frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0$

$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i$

$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$ *

$\therefore \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$ is the MLE for λ

(b) $\hat{\lambda} = \frac{5}{0.28+0.66+0.56+0.43+1.07} = \frac{5}{3}$ *

$l''(\lambda) = \frac{d}{d\lambda}\left[\frac{n}{\lambda} - \sum x_i\right]$

$= \frac{-n}{\lambda^2}$

Variance of $\hat{\lambda} \approx \frac{-1}{l''(\hat{\lambda})} = \frac{\left(\frac{5}{3}\right)^2}{5} = \frac{5}{9}$

The approximate 95% CI for λ is $\frac{5}{3} \pm 1.96\sqrt{\frac{5}{9}}$

$= 1.67 \pm 1.96 \times 0.745$

$= [0.205768\dot{9}, 3.127564]$ *

B. Risking your sanity and life, you have spent nine years gathering data on Hungarian Hornbills. You have followed these animals throughout their lives, gathering repeated measurements on individuals. These creatures are vicious and can spew extremely long flames of fire from their bills. Not much is known about Hungarian Hornbills in general, and therefore your dataset is coveted by many. The Ministry of Magical Statistics (MoMs) has tasked you to find out whether the body size of the predicts the flame reach, if that would be possible many accidents could be avoided.

Your experience with hornbills allows you to make the a priori consideration of sex – because you know that the sexes have different flame reaches. Therefore, consider an interaction between sex and body size. You run your model in R, and get the following output:

```
library(lme4)
model<-lmer(FlameReach~BodySize*Sex+(1|Individual),data=a)
summary(model)
Linear mixed model fit by REML ['lmerMod']
Formula: FlameReach ~ BodySize * Sex + (1 | Individual)
Data: a

Random effects:
    Groups     Name        Variance Std.Dev.
    Individual (Intercept) 0.5494   0.7412
    Residual               0.2119   0.4603
Number of obs: 2000, groups:  Individual, 100

Fixed effects:
                Estimate Std. Error t value
(Intercept)    4.3178    0.7569   -5.705
BodySize       0.9255    0.1013    9.134
Sex           -0.2222    0.2158   -1.234
BodySize:Sex  -0.3193    0.1328   -2.404
```

`Flamereach` is in meters, `BodySize` was standardized to a mean of 0 and a standard deviation of 1, and females are coded as 0, males as 1. Answer each of the following questions.

(i) Calculate the repeatability (in %) of the variable `FlameReach`. You do not have to give an indication of precision for the repeatability. Write out the general equation(s) of how to calculate repeatability, and a verbal explanation of each term. [30%]

(ii) Write out the two model equations for the fixed part for both sexes. You do not have to add the random part. Simplify both equations as much as possible to quantify the slope for each sex. [30%]

(iii) Write a results section dealing for your report to the MoMs, as you would a combined methods and results section in a paper. [40%]

Model Answer (Markers – Julia Schroeder (first), James Rosindell (second)):

(i) The within-individual repeatability of FlameReach is 72%. The within-individual repeatability is the proportion of total phenotypic variance (V_p) explained by between-individual effects (V_i). The total phenotypic variance is the variance explained by between-individual effects (V_i) in addition to the residual variance (V_r), which can also be considered the within-individual variance.

Eq 1 $V_p = V_i + V_r$

Eq 2 $R = V_i/V_p$

Distinction marks are given when both equations are presented and each term is verbally explained correctly. Extra points can be achieved when between- and within- terminology is not confused. For a merit, at least Eq 2 must be given correctly, and each term be verbally explained correctly. Pass answers only give correct equations but no explanations of the terms in the text, or only give correct verbal answers but no equations. To pass, at the very least, verbal or in equation, R must be described.

Continues on next page

(ii) Eq 1: $y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_{i,j}$

With $y_{i,j}$ referring to the jth repeated observation of the flame reach y of individuals i. β refers to the parameter estimates, β_0 the intercept, β_1 is the slope of body size x_1, β_2 is the effect of sex, and β_3 is the interaction effect. $e_{i,j}$ refers to the residuals, and is further partitioned into variance components but this is not shown here.

For female Hungarian horntails sex = 0, the equation is solved as follows:

Eq 2: $y_{ij} = 4.32 + 0.93 BodySize + 0 + 0 + e_{i,j}$

The intercept for females is thus 4.32, that means, an average sized female hornbill blows a flame that reaches 4.32m. With each increase in unit body size the flame reaches 93 cm further. In male Hungarian hornbills sex = 1, therefore the equation is solved as follows:

Eq 3: $y_{ij} = 4.32 + 0.93 BodySize - 0.22 - 0.32 Bodysize * Sex + e_{i,j}$

Which simplifies to

Eq 4: $y_{ij} = 4.10 + 0.06 BodySize + e_{i,j}$

The intercept for males is thus a bit smaller, an average-sized male horntail produces a flame that reaches 4.1m far. With each increase in unit body size the flame only reaches 60cm further.

A distinction answer requires Eq1, Eq 2 and the solution Eq 4, and an explanation of all terms (i, j not needed). Higher distinction answers explain i and j and explain in detail how the fixed factor works. A merit needs to present at least Eq 1 correctly (i, j not needed), and presents a mostly correct solution for both sexes, and explains all terms mostly correctly.

To obtain a pass students need to present Eq 1 and a substitution approach, with few errors that still show a general and basic degree of understanding.

(iii) I used a mixed model approach to test whether the flames of larger Hungarian hornbills reach further than those of smaller Hungarian hornbills. Due to anecdotal evidence I expect there to be a sex difference in flame reach. I modelled flame reach (measured in meter) as response variable, and body size (standardized to a mean of 0 and a standard deviation of 1) as explanatory variable. Sex (coded with females as reference variable) was added as a two-level factor. I had 10 observations on 100 individuals, and to account for these repeated measures on individuals I added the individual horntails' identity as random factor on the intercept.

I found that the interaction of body size and sex, and the main effect of body size were statistically significant. Averagely sized female horntails blew flames of about 4.30m long, and with each standard deviation of size larger, the flame would be 0.9m longer. Males had significantly shorter flames, on average 4.1m long, and these would only increase by 0.6m for each standard deviation of size. Individual identity explained a large amount of variation in flame reach. That means, individual hornbills are astonishingly consistent in the length of the flames they blow, with a repeatability of 72%.

A distinction answer deals with at least 6 of the following 7 elements correctly: fixed and random effects, describes the model, justifies the inclusion of sex, explains reference levels, units, and gives explicit parameter estimates in a biologically meaningful way. Higher marks can be given if each element is explained comprehensively. A merit lacks up to 3 of the 7 elements but describes the remaining 4 mostly correctly. A pass requires at least the description of the fixed effects parameter effect sizes in a meaningful and mostly correct way. A fail is marred with errors.

Continues on next page

Section 5: Bayesian statistics

This section has *one compulsory question* worth 60% of the total mark divided into five points (i-v). The remaining 40% will be assessed based upon your submission of the practical given to you previously in class.

You are hiking on the Pyrenees mountain range and have discovered an ancient human bone. You extracted the DNA and sequenced its genome. You were able to obtain the genotypic information for only one locus and found that the mysterious ancient human has an `AA` genotype (homozygous for Adenine). You want to make some statistical inferences on whether this sample is genetically closer to modern Spanish, French or Basque individuals.

Let's assume that your parameter of interest is $\theta = \{S, F, B\}$ representing the probability that your sample comes from a Spanish (S), French (F) or Basque (B) population, respectively.

The data are $y = \{g, f_A^{(S)}, f_A^{(F)}, f_A^{(B)}\}$ where g is the ancient genotype (so that $g = AA$) and $f_A^{(S)}$, $f_A^{(F)}$, $f_A^{(B)}$ are the known population frequencies of allele A in modern Spanish, French and Basque, respectively. Under the assumption of Hardy Weinberg Equilibrium, we know that

$$p(g = AA, f_A^{(i)} | \theta = i) = (f_A^{(i)})^2$$

for a generic population i. Note the latter equation represents the likelihood function $f(y|\theta)$.

(i) Using Bayes' law, write the equation for the posterior probability of the sample belonging to the Spanish population given the data. Assume that you have a generic prior probability $\pi(\theta)$ with known hyperparameters. Be as formal and explicit as possible. No proofs or extra calculations are required. [10%]

(ii) Let's assume that we gather the follow population allele frequencies

$$f_A^{(S)} = 0.7$$
$$f_A^{(F)} = 0.2$$
$$f_A^{(B)} = 0.1$$

and that we ask for an opinion to experts regarding a prior probability of this sample belonging to any of the tested populations. Here the opinion from our 3 experts:

(a) Dr Cobain: `It is still highly debated whether ancient humans in the Pyrenees are the ancestors of modern French, Spanish or Basque. All anthropological evidences so far are not solid enough to point towards any specific population. We have no clue!`

(b) Professor Grohl FRS: `It must be Spanish. I see no evidence why this sample should be the ancestor of any other modern population. I am 100% sure.`

(c) Mr Novoselic: `We have collected more than 1,000 ancient samples from Pyrenees so far and we were able to assign 50% of them as Spanish, 30% as French, and 20% as Basque.`

Based on this information, choose the most suitable prior distribution $\pi(\theta)$. Justify your choice. There is no right or wrong answer (although one of them is hardly acceptable) as long as it is properly justified. Formalise $\pi(\theta)$ by assigning a prior probability for each value of θ based on your choice.[10%]

(iii) Based on your chosen prior distribution, calculate the Bayes factor for model M_1 with parameter $\{\theta = S\}$ *vs.* model M_2 with parameter $\{\theta \neq S\}$. Write the equation and provide the value for the Bayes factor. Approximate any calculation as much as you wish but be reasonable (e.g. $0.82/0.19 \approx 4$ is totally fine but $0.82/0.19 \approx 3$ is not). You will not penalised for minor algebra mistakes. Provide a brief discussion on the support for M_1 or M_2.[30%]

Continues on next page

(iv) Let's assume that a new prior probability on θ is now dependent on an unknown hyperparameter τ with distribution $h(\tau)$ with $\tau = \{-1, 0, 1\}$. In other words, the prior distribution is $\pi(\theta|\tau)$ and τ can only have the discrete values of -1, 0, or -1. Using Bayes' law in hierarchical modelling, write the equation for the posterior distribution of the sample being Spanish. Be as formal and explicit as possible. No proofs or extra calculations are required. Note that both θ and τ are discrete distributions. [10%]

Model Answer (Marker – Matteo Fumagalli (1st), Tin-Yu Hui (2nd)):

(i) **[10%]**

To get the full score, one must simply write:

$$p(\theta = S|y) = \frac{f(y|\theta = S)\pi(\theta = i)}{\sum_{i\in\{S,F,B\}} f(y|\theta = i)\pi(\theta = i)}$$

with the exact specification of $\theta = S$ and the summation in the denominator.

(ii) **[10%]**

The most reasonable prior is given by expert 3 because the elicited prior is supported by a lot of previously obtained data. Option 1 could be accepted if the student feels that option 3 does not guarantee enough confidence. Option 2 is not acceptable because, as seen in class, it is not good practise to use string elicited priors that limit the parameter space.

Assuming option 3, then the prior is defined as:

$$\pi(\theta = S) = 0.50$$
$$\pi(\theta = F) = 0.30$$
$$\pi(\theta = B) = 0.20$$

Assuming option 1, then the prior is defined as:

$$\pi(\theta = S) = \frac{1}{3}$$
$$\pi(\theta = F) = \frac{1}{3}$$
$$\pi(\theta = B) = \frac{1}{3}$$

(iii) **[30%]**

Assuming option 3, then the two models have the same prior probability. Therefore,

$$BF = \frac{p(M_1|y)}{p(M_2|y)}$$

with

$$p(M_1|y) = \frac{((0.7)^2 * 0.5)}{((0.7)^2 * 0.5) + ((0.2)^2 * 0.3) + ((0.1)^2 * 0.2)} = \frac{(0.49 * 0.5)}{(0.49 * 0.5) + ((0.04 * 0.3) + (0.01 * 0.2)} = \frac{0.245}{0.245 + 0.012 + 0.002} = \frac{0.245}{0.259} \approx 0.95$$

Likewise

$$p(M_2|y) = \frac{((0.04 * 0.3) + (0.01 * 0.2)}{0.259} = \frac{0.012 + 0.002}{0.259} = \frac{0.0122}{0.259} \approx 0.05$$

Continues on next page

Then

$$BF = \frac{0.95}{0.05} \approx 20$$

A BF of 20 is indicative of positive-to-strong support for model M_1.

(iv) **[10%]**

To get the full score, one must simply write:

$$p(\theta = S|y) = \frac{\sum_{j\in\{-1,0,1\}} f(y|\theta = S)\pi(\theta = i|\tau = j)}{\sum_{j\in\{-1,0,1\}} \sum_{i\in\{S,F,B\}} f(y|\theta = i)\pi(\theta = i|\tau = j)}$$

with the exact specification of all summations.

End of paper