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, 28rd 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.

Section 1: Maths

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

A. Solve ONE of the following exercises [30%]:

(i) Solve the following integral:

$$I = \int \frac{7x-6}{x^2+x-6}dx$$

(ii) Obtain the Taylor's expansion of the following function at $x_0 = 0$ up to order three:

$$f(x) = \log\sqrt{\frac{1+x}{1-x}}$$

Model Answer (Marker – Samraat Pawar (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

$$\frac{1}{x^2+x-6} = \frac{1}{(x-2)(x+3)} = \frac{A}{(x-2)} + \frac{B}{(x+3)} = \frac{A(x+3)+B(x-2)}{(x-2)(x+3)}$$

Equating the numerator to the one we have in the integral yields the system of equations

$$A + B = 7$$
$$3A - 2B = -6$$

whose solution is $A = 8/5$ and $B = -27/5$. We can rewrite the original integral in terms of immediate integrals now:

$$I = \int \frac{7x-6}{(x-2)(x+3)}dx = \frac{8}{5}\int \frac{1}{x-2}dx - \frac{27}{5}\int \frac{1}{x+3}dx$$

$$= \frac{8}{5}\log|x-2| - \frac{27}{5}\log|x+3| + C.$$

(ii) SOLUTION: First of all, note that the function is continuous and derivable at $x_0 = 0$, we start rewriting it to simplify the operations:

$$f(x) = \frac{1}{2}\log\frac{1+x}{1-x} = \frac{1}{2}\left(\log(1+x) - \log(1-x)\right).$$

Now, we compute the derivatives:

$$\begin{aligned} f'(x) &= \frac{1}{2}\left(\frac{1}{1+x} + \frac{1}{1-x}\right) \\ f''(x) &= \frac{1}{2}\left(\frac{-1}{(1+x)^2} + \frac{1}{(1-x)^2}\right) \\ f'''(x) &= \frac{1}{2}\left(\frac{2}{(1+x)^3} + \frac{2}{(1-x)^3}\right) \end{aligned}$$

Continues on next page

which, evaluated at $x_0 = 0$ yield $f(0) = 0$, $f'(0) = 1$, $f''(0) = 0$ and $f'''(0) = 2$. Therefore, the polynomial required is

$$T_3(x) = x + \frac{2}{3!}x^3 = x + \frac{1}{3}x^3.$$

B. Solve ONE of the following sets of exercises [70%]

(i) Consider the following first order differential equation (ODE):

$$a)\ y' = \frac{y^2 - x^2}{xy}.$$

(a) Obtain the general solution $y = f(x, C)$, being C a constant. (40%)

(b) Obtain the value of C for the particular solution $y(e^2) = \sqrt{4}e^2$, where e is the Euler number (i.e. the base of the natural logarithm). (15%)

(c) Explain what the domain of the particular solution $y = f(x)$ you obtained in the previous step is. (15%)

(ii) Consider the following matrix:

$$A = \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix}$$

i. Diagonalize the matrix A, and find the matrix P such that $D = P^{-1}AP$, being D diagonal. (40%)

ii. Find the matrix P^{-1}. (15%)

iii. Verify the Cayley-Hamilton theorem, which states that any square matrix is a root of its characteristic polynomial. (15%)

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

(i) *APG comments: The exercise contains most of the topics seen in the lectures: knowledge of basic concepts and mathematical functions (domain and definition of logarithm), properties of the derivatives (derivative of the product in the change of variables) integration (to obtain the general solution of the ODE) and identification of methods for solving ODEs. Please take in consideration these points in the marks.*

(a) The equation is homogeneous, both numerator and denominator are order two, so we can start rewriting a little bit the equation dividing both numerator and denominator by x^2:

$$y' = \frac{\frac{y^2}{x^2} - 1}{\frac{y}{x}}.$$

From this expression, it is immediate to see the convenience of the classical change of variables applied to homogeneous ODEs, which is $u = y/x$. With this change we observe that $y = ux$ and, therefore, we can obtain the substitution for y' applying the derivative of the product

$$y' = \frac{dy}{dx} = \frac{d(ux)}{dx} = \frac{du}{dx}x + u\frac{dx}{dx} = u'x + u.$$

We substitute the change in the ODE and it yields

$$u'x + u = \frac{u^2 - 1}{u} \Rightarrow u'ux + u^2 = u^2 - 1 \Rightarrow u'ux = -1.$$

Continues on next page

This results in a separable ODE

$$u'u = \frac{du}{dx}u = \frac{-1}{x} \Rightarrow udu = \frac{-dx}{x}$$

Integrating both sides we obtain

$$\frac{u^2}{2} = \ln|x| + C$$

We now substitute back the original variables to obtain the general solution:

$$y^2 = 2x^2 \ln|x| + C.$$

(b) *APG: Here I want to check that they know the definition of logarithm, because we emphasized it a lot in the lectures with the definition of informational entropy and the biodiversity index, and what is meant with "particular solution".*

To obtain the particular solution, we substitute $x = e^2$ and $y = \sqrt{4}e^2$ in the general solution

$$(\sqrt{4}e^{\mathbf{2}})^2 = 2e^4 \ln|e^2| + C.$$

Using either the definition of logarithm or the properties of logarithms, we observe that $\ln|e^2| = \ln e^2 = 2\ln e = 2$, and then we easily find that $C = 0$.

(c) *APG: We ensure here that the concept of domain is understood.*

To obtain the domain of the function $y = f(x)$, we look for possible singularities. The particular solution we have found is:

$$y = \sqrt{2x^2 \ln|x|}$$

so we should be careful with the domains of the logarithm first, and then of the square root. As in the argument of the logarithm there is the absolute value of x, there are no problems with this function except in $x = 0$, where it diverges to infinity. In addition, we observe that it is multiplied by x^2 and, thus, we will have no problems with negative values. Therefore, the domain of the function would be $\mathbb{R} - \{0\}$.

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

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

We get the roots $\lambda_1 = 1$ and $\lambda_2 = -2$, which are the eigenvalues of A. To find the first eigenvector we subtract λ_1 from the diagonal of A:

$$M = \begin{pmatrix} 0 & -1 \\ 0 & -3 \end{pmatrix}$$

and we solve the homogeneous system of equations $Mx = 0$:

$$\begin{pmatrix} 0 & -1 \\ 0 & -3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0 \Longrightarrow \quad \begin{array}{rcl} -y & = & 0 \\ -3y & = & 0 \end{array}$$

from which we observe that $y = 0$ and x may have any value, so we can write the first eigenvector as $v_1 = (a,\ 0)$. Getting a simple solution, $a = 1$, it yields $v_1 = (1,\ 0)$. Proceeding similarly for $\lambda_2 = -2$ you should obtain the vector $v_2 = (a,\ 3a)$, and we pick up as particular solution $v_2 = (1,\ 3)$. Therefore, the matrix of the eigenvectors will be the matrix of change of basis

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 3 \end{pmatrix}.$$

(b) *APG: It is interesting to compute the inverse, because they should have an idea of what a cofactor matrix, adjoint matrix, transpose matrix and determinant are].* The inverse of P can be computed as $P^{-1} = \frac{1}{\det P}\mathrm{adj}(A)$, yielding

$$P^{-1} = \begin{pmatrix} 1 & -1/3 \\ 0 & 1/3 \end{pmatrix}.$$

(c) We substitute λ by A in the polynomial and we get

$$A^2 + A - 2I = \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix} + \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} = \dots$$

$$\dots = \begin{pmatrix} 1 & 1 \\ 0 & 4 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ 0 & -2 \end{pmatrix} + \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

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.

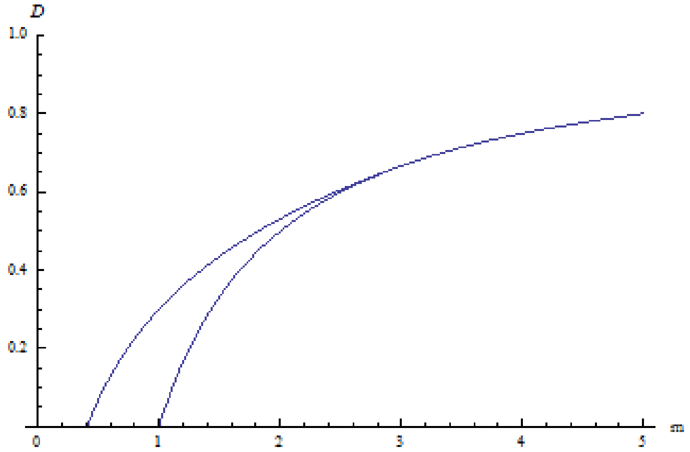
A. A model for a Levins' metapopulation with habitat destruction and in which a rescue effect operates is given by the equation:

$$\frac{dp}{dt} = mp(1 - D - p) - e_0 p e^{-ap}$$

Where p is the fraction of patches that is occupied, e_0 the extinction rate, e is the base of the natural logarithm, m the basic colonisation rate, and D is the fraction of habitat this is destroyed and not available for colonisation or occupation.

(i) Show in a graph how the colonisation and extinction rates per patch depend on p, the fraction of patches that is occupied. Indicate the values that these rates take when crossing the ordinate (i.e. when $p = 0$) [35%]

(ii) Below is a co-dimension 2 bifurcation diagram in parameters m and D for $a = 3$ and $e_0 = 1$. Bifurcations from negative equilibria are not shown. The lines form the boundaries of 3 different regions with different qualitative dynamics. What are the phase portraits in the 3 different regions? Remember that as we have only one variable in this model, the phase plane is one dimensional.[35%]



(iii) Starting from an extant metapopulation in which all habitat is available (i.e. $D = 0$) and assuming, $a = 3$ and $e_0 = 1$, what will you observe if the amount of habitat destroyed is gradually increased? Are these changes reversible?[30%]

Model Answer (Markers – Samraat Pawar (first), James Rosindell (second)):

(i) They will need to draw the functions $p = m(1 - D - p)$ and $p = e_0 e^{-ap}$. For p=0 these rates take the values $m(1 - D)$ and $e0$. These values are important because if $m(1 - D) = e_0$ there is a bifurcation, which will help figuring out question (ii)

(ii) Starting from left to right, the phase portraits are

Region 1: $- - * - - \leftarrow\leftarrow --$
Region 2: $- - * - - \leftarrow - - \circ - - \rightarrow - - * - - \leftarrow$
Region 3: $- - \circ - - \rightarrow - - * - - \leftarrow --$

Asterisks are stable equilibria. They will probably depict stable equilibria as a closed dot, and the unstable as an open dot. To work this out they will need to use the information in (i), but I don't expect details.

Continues on next page

(iii) It depends on m. Starting with $D = 0$ in region 2 above, an increase in habitat destruction will lead to a sudden (catastrophic) and irreversible drop in the occupation of the metapopulation. Starting in region 3 (m roughly larger than 3) it will lead to a drop in the occupation, but this is reversible in this case. The first transition is a catastrophic transition, the second a phase transition (if they happen to name it, not that important). I there is a lot of colonisation the rescue effect doesn't make much of a difference. Not mentioning that it depends on m gives only partial marks (say 10).

B. Before 1997 New Zealand had an immunisation programme against measles in which children were vaccinated with a first dose at age 15 months and a second dose at age 11 years. The programme used the MMR vaccine, which gives life-long protection. The vaccine had an efficacy of 90% (90% of those who received the vaccine were actually protected) and it was assumed that the vaccination scheme had a coverage of 80% (meaning that the vaccine was successfully administered to 80% of the target age class). This would theoretically result in immunisation of 92.16% of the age group over 11 (at first vaccination a fraction of $0.8 \times 0.9 = 0.72$ would be protected, at the second vaccination a further $0.28 \times 0.8 \times 09$). The basic reproductive number, R_0, was estimated to be 12.5. In 1996 mathematical modellers predicted that this immunisation scheme would be insufficient to eliminate measles, and indeed, in 1997 a measles outbreak started.

(i) Explain why you think this immunisation programme would not eliminate the disease [35%]

(ii) Describe in words what happens to number of susceptible individuals over the populations over the years [30%]

(iii) Recommend changes to the immunisation programme [35%]

Model Answer (Markers – Samraat Pawar (first), James Rosindell (second)):

(i) the fraction that needs to be immunised is 1-1/12.5 =0.92. The programme will achieve this fraction (but only just), in the population over 11. In the group under 11 about 28% will be unvaccinated, leaving the population vulnerable to a measles outbreak. If we assume that the vaccine covers 92.16% of the over 11s, and 72% of the under 11s and assume that the 1-15 months is protected by maternal antibodies, a back of an envelope calculation that there is roughly 2/15's of the population in the 72% group. That would give an R_0 of about (1-(13/15*0.92+2/15*0.72))*12.5=1.333. This is substantially over 1 and would allow for an outbreak. The calculation at the end is not essential, but if someone comes up with it they deserve some extra points. The key observation is that a part of the population is poorly vaccinated, and even though the over 11 populations is just about at a level where herd immunity would follow, the under 11 population is below this level and therefore measles outbreaks are possible. This is further enhanced if there is a lot of transmission within the under 11 group (which is likely as much of the transmission is in schools)

(ii) (20%) After an outbreak the number of susceptibles will be low and no substantial outbreaks are likely to occur. As R0 approaches 1 the number of susceptibles increases you will see bigger clusters of measles cases. Once R0 is over one large, population scale outbreaks are likely

(iii) (30%) They will need to bring the coverage rate up. Even if the second vaccination age as brought down to, say, 2 years of age, the population as a whole is still at the critical level. So vaccinating earlier will help, but on its own is unlikely to achieve herd immunity. They could mention bringing the efficacy up, or transmission down, but in practical terms that is unlikely to work. Reducing the duration of the infectious period is also unlikely to work, I don't know of drugs that could do that.

Continues on next page

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. Answer the following:

(i) Suppose there is an infinite, random-mating population with 2 alleles, A and B, with frequencies p and q, respectively, with the diploid genotypes at Hardy-Weinberg equilibrium. Suppose in AB heterozygotes the B allele is transmitted to a proportion d of gametes (where $d > 0.5$ for gene drive). In addition, BB homozygotes are lethal and produce no gametes. Construct a table to calculate the expected frequencies of allele B in the next generation. [40%]

(ii) Show that the equilibrium frequency of allele B depends on d only. [30%]

(iii) Suppose we have written a genetic drift simulator in `R`. The simulator takes three arguments: p_0, the initial allele frequency of the allele; N, the effective population size; and t, the number of generations you wish to simulate forward in time. Explain how you can estimate the variance of allele frequency due to genetic drift in the next generation, given the current p_0 and N. [30%]

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

(i)

	A	B	
Allele frequency (Now)	p	q	

	AA	AB	BB
Genotypic frequency (Now, under HW equilibrium)	p^2 $=(1-q)^2$	$2pq$ $=2(1-q)q$	q^2
selection	1	1	0 (Lethal)
frequency after selection	$\frac{(1-q)^2}{(1-q)^2+2q(1-q)}$ $=\frac{1-q}{1+q}$	$\frac{2q(1-q)}{(1-q)^2+2q(1-q)}$ $=\frac{2q}{1+q}$	0 0

For AB heterozygotes, they will produce $\begin{cases} d \text{ of allele B} \\ (1-d) \text{ of allele A} \end{cases}$ due to super-Mendelian transmittion (gene drive).

Therefore, the allele frequency of allele B at the next generation, denoted by q', is

$$q' = \frac{2q}{1+q}\, d$$

#.

(ii) At equilibrium, $q' = q$ (no change in allele frequency)

i.e. $q = \frac{2q}{1+q} d$

$\Rightarrow$ $1+q = 2d$ (Given $q \neq 0$)

$\Rightarrow$ $q = 2d-1$ ☆. (depends on d only).

(iii) - Run a large number of simulation, independently, to simulate the allele frequency in the next generation, given p_0 and N. ($t=1$ in this case)

- Record all simulated allele frequencies at time $t=1$.
- Calculate the mean, and this is your Monte Carlo estimate.

- END -

B. An engineering company is designing a bioreactor that breaks down input biomass waste (e.g. wood pulp from a paper mill) and produces ethanol that can be used for fuel. Liquid flows into the reactor containing wood pulp and liquid flows out containing the ethanol at a constant rate. The reactor contains a complex mixture of different bacteria species that together perform the metabolic steps converting cellulose into ethanol. They have asked you to produce a theoretical model to help them design an optimal system.

(i) Sketch out graphically or with equations your approach for modelling the conversion of cellulose into ethanol [35%].

(ii) What key features of the system would you tweak to improve the concentration of ethanol in the outflow [35%]?

(iii) Two species compete for an intermediate substrate (e.g. glucose): a beneficial species that converts it to ethanol and a problem species that converts it to methane, which is undesirable. What options would you explore to control or reduce the effects of the problem species [30%]?

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

(i) We covered Monod equations:

$$dC/dt = D(Q - C) - kCN/(m + C)$$

, where C is cellulose concentration in reactor, Q is input concentration, D is dilution rate, k is maximum rate of reaction, m is Michaelis-Menten constant, i.e. substrate concentration yielding half maximum rate, N is density of microbe breaking down cellulose.

$$dN/dt = ckCN/(m + C) - DN$$

, where small c is the conversion of metabolised substrate into bacterial cells

$$dE/dt = kCN/(m + C) - DE$$

, where E is ethanol concentration.

Probably occurs by multiple steps, but if equation shown assuming one step that's fine.

(ii) The steady-state concentration of E can be solved and is:

$$Q - Dm/(ck - D)$$

, where the latter term is the steady state concentration of cellulose in the reactor (this assumes 1:1 stochiometry of production). So you want to speed up the rate of reaction (lower m, higher c or K) or reduce D (distinction might notice that slowing down will increase concentration of ethanol in outflow, but production rate of ethanol would be at intermediate flow rate.)

(iii) We didn't cover this specifically, but the species reducing the concentration of cellulose most, i.e. with fastest reaction rate, will outcompete the other one. So you could engineer a faster rate strain for the ethanol production step. Credit also for talking about using phage, antibiotics etc as alternative approaches.

Overall - distinction combines light handling of equations and insights into biological features; merit can recall equations and provide basic interpretation; pass would have some general ideas but lack of linkage between theoretical ideas and the problem at hand.

Continues on next page

Section 4: Maximum Likelihood & GLMs

Please select exactly **one question** and answer it. Calculator may be required in some questions. Use the chi-square table below for critical values:

Degrees of freedom	$\chi^2_{0.95}$
1	3.84
2	5.99
3	7.81
4	9.49

A. Answer the following:

(i) Let X be a random variable following exponential distribution with rate parameter $\lambda > 0$. Given the probability density distribution $f_X(x) = \lambda e^{-\lambda x}$ and the moment generating function $M_X(t) = \frac{\lambda}{\lambda - t}$ for some t, show that $E[X] = \frac{1}{\lambda}$and $Var[X] = \frac{1}{\lambda^2}$ [30%]

(ii) After fitting a linear regression model with a slope and an intercept (and also the variance of the residuals), a student suggests to conduct a Likelihood-Ratio test (LRT) to test whether the intercept is significantly different from zero. Please describe carefully the procedures of the LRT, and how a conclusion can be drawn based on the chi-square table provided. [40%]

(iii) Explain, as precise as possible, that how you would construct the 95% confidence interval of your maximum likelihood estimates under approximate normality. You may use appropriate equations, graphics, or R commands to support your answer. Please discuss both univariate (one parameter) and multivariate (multiple parameters) cases. [30%]

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

Continues on next page

(i) $M_X(t) = \frac{\lambda}{\lambda - t}$

$$M'_X(t) = \frac{d}{dt}\left[\lambda(\lambda-t)^{-1}\right] = \lambda(-1)(\lambda-t)^{-2}(-1)$$

$$= \lambda(\lambda-t)^{-2}$$

$$E(X) = M'_X(0) = \lambda(\lambda-0)^{-2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$ ✱

$$M''_X(t) = \frac{d}{dt}\left[\lambda(\lambda-t)^{-2}\right] = \lambda(-2)(\lambda-t)^{-3}(-1)$$

$$= 2\lambda(\lambda-t)^{-3}$$

$$E(X^2) = M''_X(0) = 2\lambda(\lambda-0)^{-3} = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2}$$ ✱

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$ ✱

OR to calculate $E(X)$ and $E(X^2)$ via pdf

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \ldots$$

$$E(X^2) = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx = \ldots$$

(Integration by parts.)

(ii) - First, to fit the full model with both the slope and the intercept .

$y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

a : intercept , b: slope

- Record down the maximised log-likelihood , called l_0 .
- Then , to fit a simplified model without the intercept .

$y_i = 0 + bx_i + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Record down the maximised log-likelihood , called l_1
- Compute the likelihood-ratio test statistic , D , where

$D = 2[l_0 - l_1] \sim \chi^2_{\#\text{ of parameters dropped}}$

· If $D > \chi^2_{1, 0.95} \approx 3.84$, we tend to believe that intercept is not zero .

(iii)

- The approximate 95% CI for θ is $\hat{\theta} \pm 1.96\sqrt{Var(\theta)}$ (under approximate normality)

- $Var(\theta) \approx \left.\frac{-1}{\frac{\partial^2 l}{\partial \theta^2}}\right|_{\theta=\hat{\theta}}$, the second derivative of the log-likelihood function, evaluated at $\hat{\theta}$.

- For multivariate (multiple parameters) case, the variance-covariance structure of the estimators can be estimated by the Hessian matrix from the output of optim().
  $\hat{V}(\underline{\theta}) = -H(\hat{\underline{\theta}})^{-1}$.

- Then we can use $\hat{V}(\underline{\theta})$ to construct 95% CI for any subset of the parameters.

—END—

B. You have used data on house sparrows to see whether males were consistent in the proportion of offspring they are cuckolded with (extrali-pair offspring, EPO in the brood they care for). So, across a males lifetime, you counted the offspring a male had with their respective social partner (within-pair offspring, WPO), and those that he cares for, but did not sire himself in the same nest (EPO). You then estimated the repeatability of the number of EPO a male finds in its nest, within a year.

To do this you run a Generalized linear mixed model (GLMM). You add no fixed effects, but with male id as a random effect on the intercept. However, because the EPOs are count data, you run it with a logit link function. That also means that we have to calculate the repeatability differently from what we normally do. In Gaussian models, repeatability is the ratio of the variance explained by individual identity over the total phenotypic variance:

Eq 1: $R = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2}$

were σ_α^2 is the variance of the random effect of the MaleID and σ_ε^2 is the variance of the residual variance.

However, for the logit-link model, we calculate the link-scale repeatability RL as:

Eq 2: $R_{\text{Link}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2+(\pi^2/3\ )}$

were π is the constant pi and equals 3.14.

The repeatability on the original scale for the logit-link model needs to take into account overdispersion, and that is captured by including the fixed effects parameter estimates. The equation to estimate the repeatability on the original scale is:

Eq 3: $R_{\text{Original}} = \frac{\sigma_\alpha^2 P^2/(1+\exp(\beta_0))^2}{(\sigma_\alpha^2+\sigma_\varepsilon^2)P^2/(1+\exp(\beta_0))^2+P(1-P)}$

with Eq 4: $P = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$

The result from this is $R_{\text{Original}} = 0.10$ (95% CI = 0.08–0.12)

You get the following output from using `MCMCglmm()`:

```
model<-MCMCglmm(EPO~1, random=~MaleID, data=d, family="poisson",
nitt=500*1000, thin=10*10, burnin=3000*10, verbose=FALSE)

Iterations = 30001:499901
Thinning interval  = 100
Sample size  = 4700
DIC: 4370.051
G-structure:  ~PID

post.mean l-95% CI u-95% CI eff.samp
MaleID     2.10    1.616     3.18     2862

R-structure:  ~units

post.mean l-95% CI u-95% CI eff.samp
units    0.60   0.3673   0.8779     2356

Location effects: cbind(EPO, WPO) ~ 1

post.mean l-95% CI u-95% CI eff.samp  pMCMC

(Intercept)    -2.00   -2.428   -1.887     3678 <2e-04 ***
```

(i) Verbally describe what each of the components of the output are [25%].

(ii) Roughly estimate the link-scale repeatability of a male being cuckolded, and write down how you do that. Report both the link scale, and the original scale repeatability as you would in a paper, and interpret the original scale repeatability biologically.[40%]

Continues on next page

(iii) Verbal justify why we use the logit link function, and why this means we have to estimate two different repeatabilities. [35%]

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

(i) This is a model output from an MCMCglmm model. The response variable is the number of extra-pair offspring, EPO. There is no fixed effect present. The model ran for 500.000 iterations, with a 30.000 burn-in period. The chain was sampled at every 100th iteration. The posterior mean for the MaleID is the random effect, it is 2.10 with a 95Credibility interval of 1.62-3.18. This is estimated from 2862 samples. The residual variance is 0.57 (95% CI = 0.37–0.88). The only fixed parameter estimate we get is for the intercept, and it is -2.00 (95% CI = -2.53–1.89). Because the 95CI overlaps zero this is statistically significantly different from zero. A distinction answer would identify this as the link-scale average EPO, which translates to 0.14 EPOs on average.

(ii) Students should identify σ^2_α as MaleID parameter estimate, and σ^2_ε as the residual variance. The link-scale repeatability is thus approximately 2.1/(2.1+0.60+3.3)=2.1/6≈1/3=0.33 RLink=0.33. The link scale repeatability of a male being cuckolded is 0.33, and the original scale repeatability is 0.10, with a 95% credibility interval (some students may write confidence interval) of 0.08-0.12. This is a rather low repeatability, but it still indicates that some males are more consistent in how often their female cheats on them – approximately 10% of the variation in how often males are cuckolded is explained by differences between males.

(iii) We ran a model with a logit link function because our data consists of count data, in which we cannot assume that the residual error is normally distributed. That means that the parameter estimates are on a different scale than the original data, and thus it is difficult to interpret them in a biologically meaningful way. The back-transformation to the original scale helps with this.

Continues on next page

Section 5: Bayesian statistics

This section has **one compulsory question** worth 70% of the total mark. The remaining 30% will be assessed based upon your submission of the practical given to you previously in class.

During your latest field trip in Costa Rica you observed how brightly coloured poison dart frogs (part of the *Dendrobatidae* family) were. In fact, the brightness of their skin colouration is correlated with their toxicity. To investigate the prevalence of toxic frogs in the area under study, you collected n samples of poison dart frogs and observed that k of them have bright skin colour (and thus are toxic). We want to estimate the **population** frequency of the red colour phenotype, $f \in [0, 1]$.

(i) Assuming a generic likelihood function $p(k|f, n)$, where k is our observed data, and prior distribution $p(f)$, write the expression for the posterior distribution of f, $p(f|n, k)$. Please indicate the interval for the integration over f explicitly. [10%]

If we define the likelihood function as a Binomial distribution:

$$p(k|f, n) = \binom{n}{k} f^k (1 - f)^{n-k} \tag{1}$$

and the prior function as a Beta distribution $B(\alpha, \beta)$:

$$p(f) = \frac{1}{B(\alpha, \beta)} f^{\alpha-1} (1 - f)^{\beta-1} \tag{2}$$

then the posterior distribution of f is a Beta distribution with parameters $\alpha' = k+\alpha$ and $\beta' = n-k+\beta$.

(ii) What is the frequentist estimate of f? What is the maximum likelihood estimate of f? What is the maximum *a posteriori* mode of f using the noninformative conjugate prior $p(f) \sim B(\alpha = 1, \beta = 1)$? [15%]

(iii) Assuming we collected 100 samples and 35 of them have bright skin colour, please complete the `R` code below (fill in the '???''s) in order to generate both the exact and approximated posterior distribution of f using the informative prior $p(f) \sim B(\alpha = 2, \beta = 1)$. [20%]

```
# we evaluate our parameter f over a grid of 100 values for the whole range [0,1]
f <- seq(0, 1, ???)

# suppose we collected 100 samples and 35 of them have bright skin colour
k <- 35
n <- 100
# alpha and beta are the parameter values for the posterior Beta distribution
alpha <- ???
beta <- ???

# we now evaluate the density function to obtain the EXACT posterior distribution
y <- ???(???, shape1=alpha, shape2=beta)

# we now use Monte Carlo sampling to obtain the APPROXIMATED posterior distribution↩
    . Make a reasonable choice for the number of random samples.
y_sampled <- ???
y_sampled_distribution <- ???
```

(iv) If we use a Normal distribution as prior information, such as $p(f) = N(\mu, \sigma^2)$, we cannot derive a closed form and cannot sample directly from the posterior distribution. We can use a rejection sampling algorithm for *indirect* sampling of the posterior distribution. This algorithm requires the identification of an envelope function $g(f)$ and a constant $M > 0$ such that $p(k|f, n)p(f) < Mg(f)$. Identify both a suitable envelope function $g(f)$ and a value for M assuming that we know that the maximum density value for the posterior distribution is K. Describe what happens to the algorithm and/or the approximation if we choose $M >> K$ or $M << K$. [25%]

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

Continues on next page

(i) $p(f|n,k) = \frac{p(k|f,n)p(f)}{\int_0^1 p(k|f,n)p(f)df}$

Comment: *This is somehow similar to one of the previous year, but I'd like to give some chances to anyone who is at least aware of what discussed in class.*

(ii) They are all $\hat{p} = \frac{k}{n}$

Comment: *This is a trick. Of course they don't have to solve it but they should know that all these quantities are equal when using a flat prior. So the answer looks long but it is actually pretty quick.*

(iii)

```
f <- seq(0, 1, 0.01)
k <- 35
n <- 100
alpha <- k+2
beta <- n-k+1
y <- dbeta(f, shape1=alpha, shape2=beta)
y_sampled <- rbeta(n=1e5,  shape1=alpha, shape2=beta)
y_sampled_distribution <- hist(y_sampled)
```

Comment: *We did these examples many times during the class so it should be straightforward, but I want to check whether they understood the concept of Monte Carlo sampling.*

(iv) If $g(f) \sim U(0,1)$ and $M = K$ then the algorithm is:

(a) Generate $f_i \sim U(0,1)$,

(b) Generate $U \sim U(0,1)$,

(c) If $MUg(f_i) <= p(k|f,n)p(f)$ accept f_i otherwise reject f_i.

Repeat this procedure until N samples are obtained.

The members of this sample will be random variables from the posterior distribution of f.

If $M >> K$ then the algorithm is inefficient as the acceptance rate is low, if $M << K$ the the approximation will be poor.

Comment: *This requires some thinking but it is something we covered a lot.*

End of paper