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, 23rd March 2016, 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. All sections allow you to choose between two questions, answering one. Read instructions carefully at the head of each section.

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 II

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

A. Preston (1962) and MacArthur and Wilson (1963) investigated the effect of area on species diversity in oceanic islands. It is assumed that species can immigrate to an island from a species pool of size P and that species on the island can go extinct. We denote the immigration rate by $I(S)$ and the extinction rate by $E(S)$, where S is the number of species on the island. Then the change in species diversity over time is

$$\frac{dS}{dt} = I(S) - E(S) \tag{1}$$

For a fixed island, the simplest functional forms for $I(S)$ and $E(S)$ are

$$I(S) = c(1 - \frac{S}{P}) \tag{2}$$

$$E(S) = \frac{mS}{P} \tag{3}$$

where c, m and P are positive constants.

(i) Find the equilibrium species diversity $\hat{S}$ of eqn 1 with $I(S)$ and $E(S)$ given in 2 and 3. [20%]

(ii) It is reasonable to assume that the extinction rate is a decreasing function of island size. That is, we assume that if A denotes the area of the island, then m is a function of A with $\frac{dm}{dA} < 0$. Furthermore, we assume that the immigration rate I does not depend on A. Use these assumptions to investigate how the equilibrium species diversity changes with island size. [20%]

(iii) Assume that $S(0) = S_0$. Solve eqn 1 with $I(S)$ and $E(S)$ given in 2 and 3. [40%]

(iv) Assume that $S(0) = 0$. That is, the island is initially devoid of species. One then defines the time constant T as the time required for S to attain 63.2% of its equilibrium value i.e. it is the solution of

$$S(T) = (1 - e^{-1})\hat{S}$$

Find an explicit expression for T in terms of P , c, and m. [20%]

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

Please see attachment

B. You are planning to conduct a field study to determine the relative abundance of a certain variety of a given species, i.e. the proportion with which this variety is present among the total population. How large a sample size would you need to estimate this proportion within 0.01 of the true value with probability at least 0.95?

Hint: Use the central limit theorem. You may need some of the following Φ values (the cumulative distribution function of the standard normal distribution):

$\Phi(1.28) = 0.9$
$\Phi(1.44) = 0.925$
$\Phi(1.65) = 0.95$
$\Phi(1.96) = 0.975$

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

Please see attachment

Continues on next page

Section 2: Dynamical Models in Ecology and Evolution

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

A. Sheep were first introduced in Tasmania in 1810. Over the period 1814–1864 their numbers increased as follows:

Year	No. of sheep (thousands)
1814	125
1824	275
1834	830
1844	1200
1854	1750
1864	1650

Based upon these data, answer the following:

(i) How could you evaluate whether or not the growth of the sheep population is density dependent? Outline/explain the possible methods. [40%]

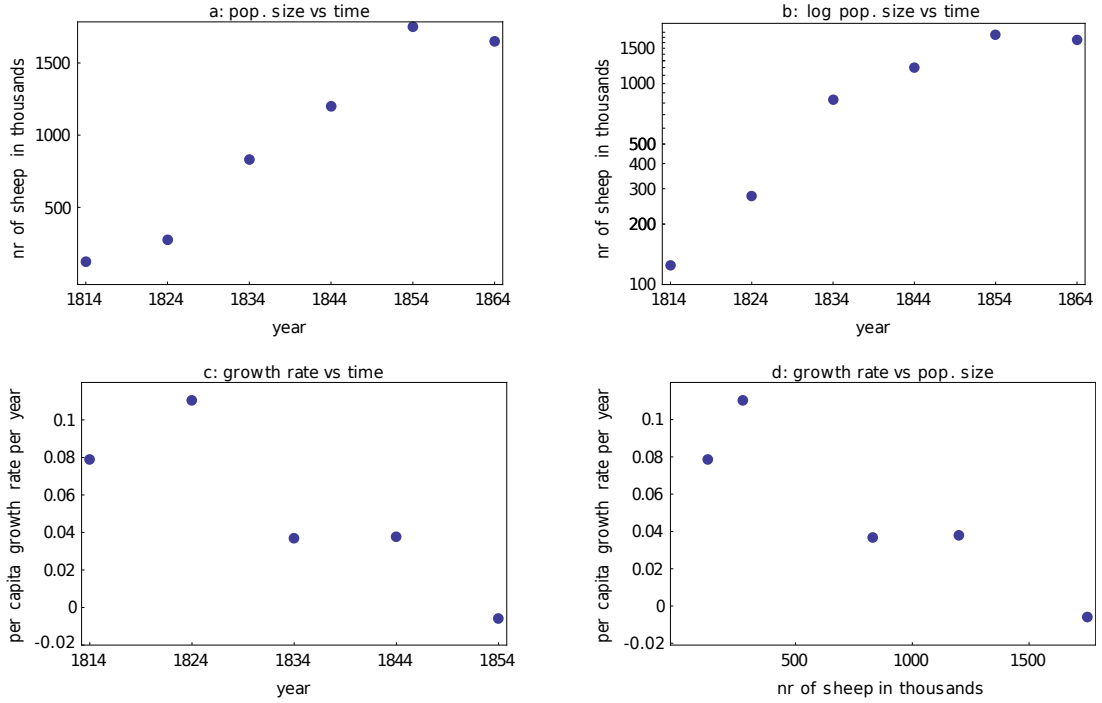
(ii) Based on the information in the table, discuss whether the growth of the sheep population is likely to be density dependent or not. [30%]

(iii) Discuss what mechanism could have caused the pattern of population growth. [30%]

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

(i) The obvious first thing to do is to plot the data and see what is looks like. Better, or next, plot the data on a semi-logarithmic plot. This will reveal whether or not the growth is exponential.

The Figure below shows: (a) the no of sheep vs time, (b) number of sheep on a logarithmic scale versus time, (c) gives the per capita growth rate versus time and (d) the per capita growth rate versus the size of the sheep population.



A next good step is to calculate the per capita growth rates between time steps ($\frac{1}{T}ln(\frac{N(t+T)}{N(t)})$) and see if these are constant over time. If they are not constant the growth is not exponential and therefore to some degree density dependent. The most illuminating plot is to plot the growth rate against population size, $N(t)$. This will reveal the pattern of density dependence. A decreasing

Continues on next page

strait line would be the pattern that corresponds to logistic growth. Even more sophisticated is to formulate a stochastic density independent growth model and several density independent model and apply Bayesian reasoning and a model selection argument. I did not cover that in my lectures, but a first class answer could mention this. Note though that to do this there need to be and some detail what the density dependent model could be (eg., logistic) and how the deviation from the model is evaluated (eg., it could be considered normal). Many choices are possible here.

(ii) Graph a shows that the growth tails off. Graph b shows that the data points do not lie on a straight line, making exponential (density independent) growth unlikely. Graph c confirms that the growth rates go down. Graph d unequivocally shows density dependence. If you would fit a line through these points there is a clear downward trend. A very good answer will tell you that this all hinges on the last data point though. It would be cautious to try to find more data to make a definitive statement. Based on these data the growth appears to be density dependent and a logistic growth model would be a reasonable choice

(iii) Not a lot, but you know the feedback is negative (density dependence), based on the graph d it looks like the logistic model would be a good fit, suggesting that there is effectively a single dominant factor regulating the population. If I would have to hazard a guess it is the amount for land for grazing that is limiting)

Graph d above possibly shows the typical pattern of logistic growth, caused normally be depletion of a resource (although there could be more, given the data one resource would do). In the first two decades the growth is positive and there is no sign of a reduction in the growth rate, after that the growth rate comes down, indicates a depletion of resources. Although we cannot tell from the data what the resource is, if I were to hazard a guess it is most likely running out of land to graze on.

B. Historical measles data often show a biannual pattern, with measles outbreaks in every second year. Between 1928 and 1935 triennial cycles were observed in the incidence of measles in Baltimore.

(i) Explain how triennial cycles can occur in measles epidemics. [40%]

(ii) Discuss how the periodicity of measles varies with the reproductive number (R_0) and the strength of forcing. [40%]

(iii) Based on the occurrence of triennial cycles, what can you say about the value of the reproductive number (R_0) and the strength of forcing in Baltimore in the 1930s? [20%]

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

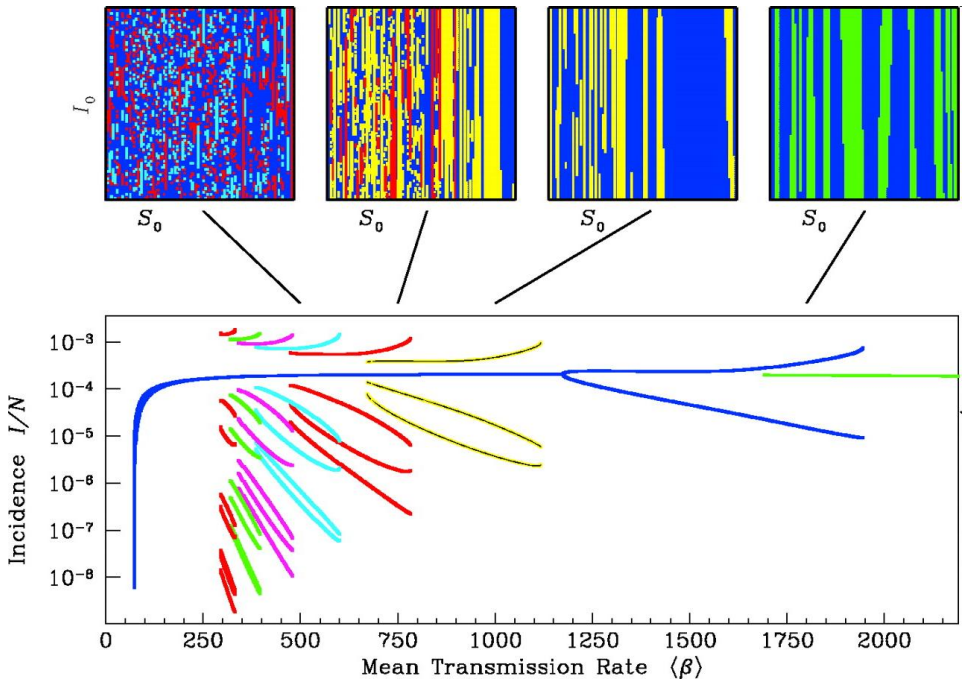
(i) **Seasonal forcing** can cause measles epidemics to skip on or more years and lead to regular cycles. This is a resonance phenomenon that occurs if the natural cycle time of the epidemic is close to an integer number of the forcing period. In simple terms, once on outbreak occurs, it takes some time for the number of susceptibles to return to levels where the next outbreak can occur. While the number of susceptibles is low, there are no signicificant outbreaks. If the number of susceptibles has increased to sufficiently high numbers (through the birth of new susceptibles) the next outbreak can occur. If the reproductive number is high, which it usually is, the pattern settles on a biannual cycle. **So typically, measles have biannual cycles**. For lower reproductive numbers the natural fluctuation time of the epidemic is lower (for instance because the transmission is lower), and the next integer number the pattern can settle on is 3.

(ii) **The lower the R_0 the higher the cycle number**. The answer can only be one centred around the Arnold tongues. The regions where the cycles takes on a particular value are organised as Arnold tongues. The tip of the tongue is close to the point where the unforced model has exactly that cycle time. Approximate formula is

$$\frac{2\pi}{\sqrt{\mu\gamma(R_0-1)}}$$

Continues on next page

(But I don't really expect that they reproduce the formula, the key points are that the period will go up as R_0 decreases and that this the windows with the same period are confined to parts of the parameter space). Earn et al. do a beautiful figure:



(iii) **NEW ANSWSER**: **Not a lot as it could be either.** Chances are it is the R0, as the seasons remain the seasons. Although not clear from the Earn figure, for different amounts of forcing the tongues for 2 and 3 start to overlap. If they really recall and understand one could argue that there is more forcing, and a slightly lower R0. It is hard to give a definitive indicative answer but there is a lot of scope for intelligent comments.

OLD ANSWSER: **The R_0 must be lower**. If we take mu approx. 1/75 and gamma about 50 (corresponding to a av. lifetime of 75 years) and an infectious period of a week, this is an R_0 of about 7-8. However, as the Arnold tongues widen with increase forcing this could be easily between 6 and 10. But again, the value is not so important. The triennial cycles suggests that for some reason the R_0 was lower.

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) Drive is one of five processes affecting allele frequencies within a population. What are the other four? [10%]

(ii) Construct a model of allele frequency changes due to drive along the following lines. 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 frequencies. Suppose in AB heterozygotes the B allele is transmitted to a proportion d of gametes ($d = 0.5$ being Mendelian inheritance with no drive). What is the expected frequency of the alleles in the next generation? If $p = 0.8$, $q = 0.2$, and $d = 0.95$, what are the expected frequencies in the next generation? What is the long-term equilibrium frequency of the two alleles (i.e., frequencies to which the population will tend many generations in the future)? [35%]

(iii) Suppose, in addition, BB homozygotes are embryonic lethal. Find an expression for the change in the frequency of the B allele from one generation to the next. Find the equilibrium frequency. What fraction of the population are BB homozygotes and die at this equilibrium? [35%]

(iv) Explain how the process of drive might be used for public health. [20%]

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

(i) Mutation, immigration, selection and (Random genetic) drift

(ii) $q' = q2 + 2dpq = q(q + 2dp)$. If $p = 0.8$, $q = 0.2$, $d = 0.9$, $q' = 0.344$, $p' = 0.656$. The long-term equilibrium frequencies are $q = 1$ and $p = 0$.

(iii) If BB are homozygous lethal, then

	AA	**AB**	**BB**
Zygotic frequency	p^2	$2pq$	q^2
Survival	1	1	0
Adult frequency	$p^2/(p^2 + 2pq)$	$2pq/(2pq + q^2)$	0
A:B gamete frequency	1:0	1-d : d	0:1

$q' = 2pqd/(p2 + 2pq) = 2qd/(1 + q)$

$\Delta q = q' - q = 2qd/(1 + q) - q$

At equilibrium, $\Delta q = 2qd/(1 + q) - q = 0$, therefore $q = 2d - 1$ (ignoring the trivial equilibrium $q = 0$).

Equilibrium frequency of BB homozygotes $= q^2 = (2d - 1)^2$

(iv) A driving gene can be used to impose a genetic load on mosquitoes that transmit disease, reducing the numbers of those mosquitoes, and so reducing disease transmission.

B. Messrs Barraclough & Co, the famous red wine manufacturers, have invented a new process for producing extra strong claret. By genetically engineering a super' strain of yeast, they pump grape juice through a continuous flow-through system, and produce an extremely high alcohol wine for the mass market.

Assume that the dynamics of conversion from glucose to ethanol and of the growth of the yeast in the

Continues on next page

system are specified by the following equations:

$$\frac{dS_1}{dt} = D\,(Q_1 - S_1) - \frac{aS_1}{(b + S_1)}N$$
$$\frac{dS_2}{dt} = \frac{aS_1}{(b + S_1)}N - DS_2$$
$$\frac{dN}{dt} = c\frac{aS1}{(b + S_1)}N - DN$$

where,
S_1 is the concentration of glucose (moles per litre) in the flow-through system
Q_1 is the concentration of glucose in the input grape juice
S_2 is the concentration of ethanol (moles per litre)
N is the density of yeast in cells per litre
D is the rate of flow through of juice (litres per second)
a is a rate parameter (sometimes called V_{max} in the microbial literature)
b is the Michaelis-Menten constant
c is the number of cells produced per mole of glucose metabolised per litre

(i) Explain each component term in the three equations in words and/or graphically. [10%]

(ii) What are the units of the constants a and b? [10%]

(iii) What is the concentration of alcohol that is produced at steady-state? How would you tweak the parameters to increase the concentration? [40%]

(iv) Imagine that parameters a, b, and c are now variables that can evolve in the yeast. Describe how you would approach including evolution into the model, and make verbal predictions for how the yeast population would evolve. [40%]

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

(i) Equation 1 has inflow of glucose ($-DQ$), outflow of glucose ($-DS1$), metabolism of glucose, which is Michaelis-Menten kinetics, rate of reaction saturated as concentration of glucose increases: maximum rate is a and concentration at half maximum rate is b. Equation 2 ethanol is produced by metabolism (1:1 stochiometry) and lost in outflow ($-DS2$). Glucose metabolism converted to cell division (proportionality constant c) and cells lost in outflow.

(ii) b is in moles per litre – it's the concentration at which rate is half, hence same units as concentration S_1. a is in moles per cell per second. The$S_1$and S_2 units cancel, hence multiply N by a gives units of moles per litre per second, units for dS_1/dt

(iii) First solve for S_1 then easy substitution to get S_1, which is what we want. S_1comes from equation 3, note that at steady-state, $dN/dt = 0$, then N's cancel and rearrangement gives $S_1 = \frac{Db}{ac-D}$

We didn't do this in class, but we talked about the concept of how you'd do it and they knew had to solve steady-states for equations like this.

Then can add equations 1 and 2, the horrible bit cancels to give

$$D\,(Q_1 - S_1) - DS_2 = 0$$

i.e., $S_2 = Q_1 - S_1 = Q_1 - \frac{Db}{ac-D}$

Now can say alcohol concentration increases linearly with concentration of glucose in the grape juice, decreases linearly with the Michaelis-Menten parameter b, increases as the product ac increases (non-linearly up to a maximum of Q_1) and concentration decreases nonlinearly as D increases (not enough time to do the conversion fully before juice flows out). These predictions assume the yeast population can grow fast enough not to be just flushed out: $D > acQ_1/(Q_1 + b)$ not expecting them to derive these conditions but certainly credit given if they do!

Continues on next page

(iv) Can add in additional equation of a form like:

$$\frac{da}{dt} = m\frac{d\left((1/N)(dN/dt)\right)}{da}$$

which is mutation rate times the gradient of the per capita rate of population increase (i.e. fitness) as parameter a changes. Similar equations could be specified for b and c. We talked in the class about kind of trade-offs, and distinction-level students would attempt to simplify by talking about trade-offs, e.g. can't increase a, b and c, but increase in one decreases another. Without trade-offs, should increase a and c, decrease b (both increase rate of population increase). But there could be feedbacks as this in turn will affect the substrate steady-state concentrations. I will give credit for any sensible comments here, not expecting much in terms of actual equations.

Continues on next page

Section 4: Maximum Likelihood

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 binomial random variable with probability mass function

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Show that $\sum_{k=0}^{n} f_X(k) = 1$. [30%]

(ii) In 30 independent coin tosses, 21 of them were head. Let p be the probability of getting a head, and the likelihood function of p is

$$L(p) = \binom{n}{y} p^y (1-p)^{n-y}$$

where n is the number of independent trials and y is the number of heads observed.

Please perform a likelihood ratio test for $H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$ at 5% significance level. [40%]

(iii) Describe, as precisely as possible, that how you can obtain 95% confidence interval (and joint confidence region for **two parameters**) from the log-likelihood function. You may use appropriate equations or graphs to support your answer. [30%]

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

Please see attached

B. Answer the following:

(i) Let $X \sim N(\mu, \sigma^2)$ and the associated moment generating function is $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. Show that $E[X] = \mu$ and $E[X^2] = \mu^2 + \sigma^2$. [40%]

(ii) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed samples from $Poisson(\lambda)$. Given the probability mass function of an Poisson random variable is $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, please find the MLE for λ. [30%]

(iii) A CMEE student is trying to implement MLE in R. She writes her own log-likelihood function log.like which contains two parameters, plus an input dataset dat. She then uses optim() to maximise the log-likelihood function. She types the following command into her R console:

```
> optim(par=c(100,0.5), fn=log.like, method= L-BFGS-B , lower=c(1,0),
upper=c(9999,1), dat=dat, control=list(fnscale=-1), hessian=T
```

Please describe, as precisely as possible, *each component of the input and output screen.* [30%]

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

Please see attached

Continues on next page

Section 5: GLMs & Bayesian stats

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

A. A colleague has conducted an observational study focused on a species of beetle present on all twenty islands of an archipelago. On each island, she counted the number of individuals she saw in one hour of observation, and repeated this at ten different spots on the island. She also measured the distance to the nearest food resource, in meters, for each spot where she had counted beetles. She then determined whether a major predator of the beetle was present or absent on each island.

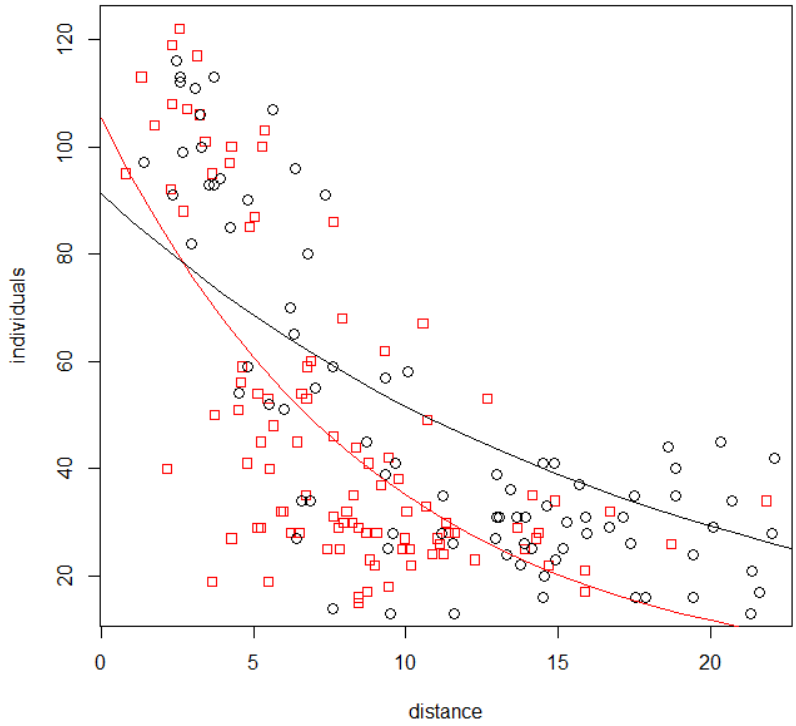
```
> str(beetle_counts)
'data.frame':   200 obs. of  4 variables:
 $ predator   : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
 $ island     : Factor w/ 20 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ distance   : num  14.26 6.49 7.97 5.14 7.41 ...
 $ individuals: int  27 28 30 29 25 17 22 23 25 28 ...
```

She hypothesizes that there will be fewer beetles farther away from food resources in the absence of predators. She also hypothesizes that this effect of distance from food in reducing the number of beetles observed will be even stronger when predators are present on the island. She has fit a generalized linear model in order to test this:

```
> mymodel<-glm(individuals ~ predator * distance, family=poisson(link=log), data=beetle_counts)
> summary(mymodel)

Call:
glm(formula = individuals ~ predator * distance, family = poisson(link = log),
    data = beetle_counts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3047  -2.0802  -0.2908   2.1894   6.7238

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.655983   0.030484 152.737  < 2e-16 ***
predatorYES        -0.142953   0.040329  -3.545 0.000393 ***
distance           -0.109915   0.004181 -26.291  < 2e-16 ***
predatorYES:distance  0.053042   0.004683  11.327  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3221.6  on 199  degrees of freedom
Residual deviance: 1633.3  on 196  degrees of freedom
AIC: 2746.4
```

The plot below shows the raw data along with predicted values from the model across a range of distance from food source:

Continues on next page



(i) Based on the model, show mathematically how you would calculate the predicted value for an island with predators, at a site 5 meters from a food resource. [30%]

(ii) Which color of line represents predictions for "predator present" and which color represents "predator absent"? [10%]

(iii) If we decided to trust this model, what can you say about your colleague's hypotheses? (30%)

(iv) The model shows symptoms of overdispersionhow can you tell this? Instead of using a "quasipoisson" error distribution to deal with the issue, what other (and better) approach could you take that incorporates an aspect of the sampling neglected by your colleague? Explain in detail why you would use this approach and also how you would do this, including either R code or pseudocode. [30%]

Model Answer (Marker – Brian Hollis (1st), Julia Schroeder? (2nd)):

(i) Calculate the model prediction for observed beetles by inserting the coefficients and the values given into the model formula.

y = B1 + B2x + B3x ...

y = 4.656 + -0.1423(1) + -0.110(5) + 0.053(5) = 4.229

(the final numeric answer is irrelevant if they are plugging the right things in before).

Next, and this should be half of the credit, they should use the inverse link function to get the predicted value onto the real scale.

exp(4.229) = 68

(ii) Black = predator present; Red = predator absent

(iii) The model indicates a significant effect of distance, so the first part of her hypothesis (there will be fewer beetles as you move farther away from a food resource, in the absence of predators) is supported. However, the second part of her hypothesis was that this effect would be stronger when predators are present. In fact, the opposite is seen by the positive interaction parameter estimate, which is significant. So, when predators are present the effect of distance from food on how many beetles are actually seen is not magnified, but instead reduced.

Continues on next page

(iv) The most obvious indication of overdispersion is the ratio of residual deviance to degrees of freedom, and pointing this out would be sufficient. They might also make a similar argument directly from the plot (the spread of points around the predicted lines is more than would be expected with poisson error, and this is OK if sufficiently clear.

A generalized linear mixed model should be used, in order to account for sampling at the island level. They should show they understand this conceptually (some islands may differ for unknown reasons, and since multiple sites are sampled on each island this needs to be controlled for). They might frame this differently, e.g. as being about how sites are essentially pseudoreplicated in the original model, which does not control for this shared island effect. Some responses might suggest collapsing the data so that there is only one measure per island, which would be OK but complete answers must go into mixed modeling.

To specify this GLMM, they might come up with the following to fit a random intercept using lme4:

```
mymodel<-glmer(individuals ~ predator * distance + (1 | island), family=poisson, data=beetle counts)
```

Or, if they fit a random slopes model:

```
mymodel<-glmer(individuals ~ predator * distance + (1 + distance | island), family=poisson, data=beetle_counts)
```

But there are other possibilities; the syntax is irrelevant, it's only important that they properly indicate what would be treated as fixed and what would be treated as random.

B. Answer the following

(i) State Bayes' theorem, naming all the symbols you use, and describe, in general terms, how Bayes' theorem can be used to perform Bayesian inference. [30%]

(ii) Explain the difference between an informative and a non-informative prior. Give an example for each. [20%]

(iii) In 2010 Ott and Rogers found that "Gregarious desert locusts have substantially larger brains compared with the solitarious phase". Given a set of brain size data for each of the two types of locusts, how would you use Bayesian inference to substantiate the authors' claim? [50%]

Your answer should include:

i. A statistical model suitable for coding in JAGS, including suitable priors; motivate your choice!

ii. The variable(s) supplied as data

iii. The variable(s) you would monitor in JAGS, and

iv. How you would interpret the results so as to validate the claim (or not). Don't worry too much about JAGS syntax, nor the actual values of the parameters of the priors you use.

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

(i) The posterior density of a random variable X given the evidence $Y = y$ is proportional to the product of the sampling density $f_{Y|X}(y|x)$, i.e. the conditional density of Y given the cause $X = x$, with the prior density of X,

$p_X(x) : p_{X|Y}(x|y) = C f_{Y|X}(y|x) p_X(x)$,

where C is a normalisation constant equal to $\frac{1}{\int f_{Y|X}(y|x) p_X(x) dx}$.

The value y of Y represents the observed data (evidence) one wishes to analyse, so as to infer knowledge about the variable X. The sampling density is provided by the statistical model of this problem, which models the statistics of the evidence given the various values of X. The

Continues on next page

prior density is a model for the prior knowledge one has about X, and the posterior density then models the combined information including prior knowledge and the evidence.

(ii) An informative prior is a model for prior knowledge incorporating actual knowledge elicited from, e.g. an expert; example: Beta(a,b) where a and b are such that the mode of the density coincides with the expert's best guess of the true value of X, and 95% of the probability mass is below (or above) the expert's guess of the absolute maximal (minimal) value.

A non-informative prior is a model for complete ignorance about the subject; for example: beta(1,1), corresponding to an empty experiment with 0 trials; a flat (uniform) distribution.

(iii) Let the data of brain sizes be contained in two vectors y1 (gregarious) and y2 (solitarious) of dimension N1 and N2, resp.

We assume that brain size in each group is normally distributed with unknown mean and variance. For the prior of the mean we use a normal distribution (as this is the conjugate distribution), for the prior of the precision a gamma distribution. These priors should be non-informative as we want to remain unbiased and only look at the available measurement data.

```
model {
for(i in 1:N1) {
y1[i] ~ dnorm(mu1,tau1)
}
for(i in 1:N2) {
y2[i] ~ dnorm(mu2,tau2)
}
tau1 ~ dgamma(1,0.01)
tau2 ~ dgamma(1,0.01)
mu1 ~ dnorm( mu0 , tau0)
mu2 ~ dnorm( mu0 , tau0)
Delta <- mu1-mu2
}
```

Here, `mu0` is an average brain size which will presumably be a few cubic millimeters, and `tau0` the prior precision, which should be as low as possible.

Supplied data: the vectors y1, y2.

Monitor Delta to see if it significantly deviates from 0, and check if 0 lies in the 95% PI or not.

End of paper

(1) a) equilibrium means $dS/dt = 0$

hence $I(S) = E(S)$

$$\Leftrightarrow c - c\frac{S}{P} = m\frac{S}{P}$$

$$\Leftrightarrow c = (m+c)\frac{S}{P}$$

$$\Leftrightarrow \boxed{\hat{S} = \frac{c}{m+c}P}$$

b) assuming m decreases with A
we find that $\hat{S}$ increases with A

c) $$\frac{dS}{dt} = c - (m+c)\frac{S}{P}$$

one possible method: substitute

$$S = x + \frac{c}{m+c}P$$

$$\downarrow$$

$$\frac{dx}{dt} = -(m+c)\frac{x}{P}$$

$$\downarrow$$

$$x = A\exp\left(-\frac{m+c}{P}t\right)$$

$$\downarrow$$

$$S = A\exp\left(-\frac{m+c}{P}t\right) + \frac{c}{m+c}P$$

$$S(0) = A + \frac{c}{m+c}P = S_0$$

$$\Rightarrow A = S_0 - \frac{c}{m+c}P$$

$$\Rightarrow S(t) = \left(S_0 - \frac{c}{m+c}P\right)\exp\left(-\frac{m+c}{P}t\right) + \frac{c}{m+c}P$$

d) $S_0 = 0$

$$\Rightarrow S(t) = \underbrace{\frac{c}{m+c} P}_{\bar{S}} \left(1 - \exp\left(-\frac{m+c}{P} t\right)\right)$$

So $S(T) = (1 - \exp(-1))\,\bar{S}$

$$\Leftrightarrow \quad \frac{m+c}{P} T = 1$$

$$\Leftrightarrow \quad T = \frac{P}{m+c} \qquad \checkmark$$

e) T also increases with A

(2) Let S_n be the # species of that variety in a sample of n.

Let p be the population abundance.

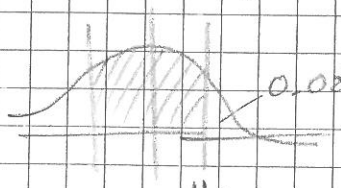
We want n s.t. $\Pr\left(\left|\frac{S_n}{n} - p\right| \le 0.01\right) \ge 0.95$

By CLT: S_n is approx normal with $\mu = np$ and variance $np(1-p)$

$$\Rightarrow y = \frac{S_n - np}{\sqrt{np(1-p)}} \text{ approx std normal}$$

now $\left|\frac{S_n}{n} - p\right| \le 0.01 \iff |y| \le \frac{0.01\,n}{\sqrt{np(1-p)}} = u$

We also have $\Pr_n(|y| \le u) \ge 0.95$



$\Updownarrow$

$2\Phi(u) - 1 \ge 0.95$

$\Updownarrow$

$\Phi(u) \ge 0.975$

$\Updownarrow$ table

$u \ge 1.96$

So, $\frac{0.01\,n}{\sqrt{n}\sqrt{p(1-p)}} \ge 1.96$

$\iff \sqrt{n} \ge \frac{1.96\sqrt{p(1-p)}}{0.01}$

Sadly: p unknown, but worst case is $p = \frac{1}{2}$

$$\Rightarrow n \ge \left(\frac{1.96}{0.01} \cdot \frac{1}{2}\right)^2 = 9604$$

Q1 (i) Consider $(a+b)^n$ for any a, b.

$$(a+b)^n = a^n + C_1^n a^{n-1} b^1 + \cdots + C_{n-1}^n a^1 b^{n-1} + b^n \quad \text{(binomial expansion)}$$

$$= \sum_{k=0}^{n} C_k^n a^k b^{n-k}$$

put $a = p$, $b = 1-p$,

$$\text{LHS} = [p + (1-p)]^n = 1^n = 1$$

$$\text{RHS} = \sum_{k=0}^{n} f_x(k)$$

Result follows.

Q1 (ii) $n = 30$, $y = 21$, under H_0, the maximised likelihood value

$$= L(0.5) = C_{21}^{30} 0.5^{21} (1-0.5)^9 = C_{21}^{30} 0.5^{30}$$

Under H_1, the likelihood is maximised when $p = \frac{21}{30} = 0.7$ (no need to prove it). Hence the maximised likelihood value

$$= L(0.7) = C_{21}^{30} 0.7^{21} (1-0.7)^9 = C_{21}^{30} 0.7^{21} 0.3^9$$

The Likelihood-Ratio test (LRT) statistic

$$= D = 2 \times [\ln(L(0.7)) - \ln(L(0.5))]$$

$$= 2 \times [21 \ln 0.7 + 9 \ln 0.3 - 30 \ln 0.5]$$

$$= 2 \times 2.468486 = 4.936972 > \chi^2_{1, 0.95}$$

$\therefore$ H_0 is rejected. We tend to believe that $p \neq 0.5$ (the coin is not fair)

P.T.O.

Q1 (iii) For univariate case, the 95% CI for parameter (θ) is the region of θ s.t. the log-likelihood remains within 1.92 (or 2) units from its maximum. (or $\frac{1}{2}\chi^2_{1,0.95}$)

For bivariate case, the 95% joint confidence region for parameters $\underline{\theta}$ is the collection of parameter values for which the log-likelihood decreases by no more than $\frac{1}{2}\chi^2_{2,0.95} = \frac{1}{2} \times 5.99 \approx 3$ units from its maximum.

Other reasonable answers, such as the use of approximate normality to find CI (Wald CI), are also accepted, although not required.

Q2 (i) $\begin{cases} E(X) = M_X'(0) \\ E(X^2) = M_X''(0) \end{cases}$

Some credits will be given if the student is able to describe / demonstrate how moments can be obtained through differentiating the mgf.

$$M'_X(t) = \frac{d}{dt} M_X(t) = \left(e^{\mu t + \frac{1}{2}\sigma^2 t^2}\right)\left(\mu + \tfrac{1}{2}\sigma^2(2t)\right)$$

$$E(X) = M'_X(0) = e^0(\mu + \sigma^2(0)) = e^0(\mu) = \mu \quad \#$$

$$M''_X(t) = \frac{d}{dt} M'_X(t) = \frac{d}{dt}\left(\mu e^{\mu t + \frac{1}{2}\sigma^2 t^2} + \sigma^2 t\, e^{\mu t + \frac{1}{2}\sigma^2 t^2}\right)$$

$$= \mu\left[\left(e^{\mu t + \frac{1}{2}\sigma^2 t^2}\right)\left(\mu + \tfrac{1}{2}\sigma^2(2t)\right)\right] + \sigma^2 t\left[\left(e^{\mu t + \frac{1}{2}\sigma^2 t^2}\right)(\mu + \sigma^2 t)\right] + \sigma^2 e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

$$E(X^2) = M''_X(0) = \mu\left[(e^0)(\mu + 0)\right] + \sigma^2(0)\left[e^0(\mu + 0)\right] + e^0 \cdot \sigma^2$$

$$= \mu(\mu) + 0 + \sigma^2 = \mu^2 + \sigma^2 \quad \#$$

Q2 (ii) $L(\lambda) = f(x_1, x_2, \cdots, x_n | \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$ ($\because$ iid Poisson(λ))

$$= (e^{-n\lambda}) \frac{\lambda^{x_1 + x_2 + \cdots + x_n}}{x_1!\, x_2! \cdots x_n!}$$

$$\ell(\lambda) = -n\lambda + (x_1 + x_2 + \cdots + x_n)\ln\lambda - \ln(x_1!\, x_2! \cdots x_n!)$$

$$= -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda - \text{constant}$$

$$\ell'(\lambda) = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} - 0$$

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

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

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

PTO.

Q2 (iii) optim(...) is the function to optimise a given functi

par = C(100, 0.5) the initial parameter values to be optimised over.

fn = log.like the functin to be optimised

method = "BL-BFGS-B" the algorithm to be used to optimise the function

lower = c(,) the lower bound of the parameter space

upper = c(,) the upper bound of the parameter space

dat = dat the extra argument / dataset to be passed to log.like

Control = list (fnscale = -1) fnscale=-1 means to maximise a functin

hessian = T to return the hessian matrix.