BE 175 Midterm, Winter 2022

Question 1 (10 points)

You are working for a biotechnology company developing a monoclonal antibody. As part of the process for isolating individual antibody-secreting B cells from an infected individual, you deposit B cells at a limiting dilution into a 96-well plate (where they are dilute enough that most wells are empty). The number of cells in each well follows a Poisson distribution:

$$p(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

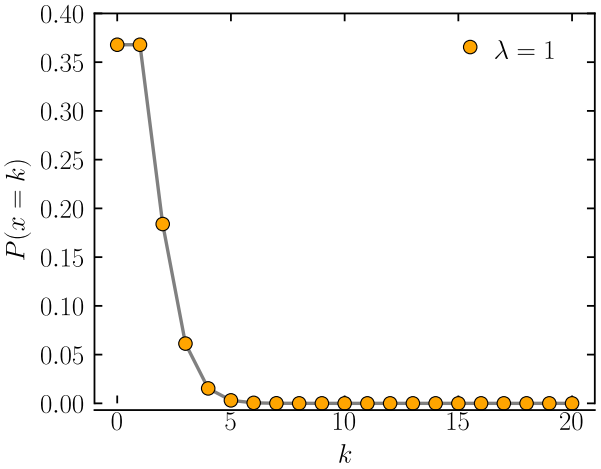
where k is the number of cells in a given well, and λ is the average number of cells per well. You want to make sure that your limiting dilution process is indeed leading to wells with mostly single cells.

a) Provide an expression for the probability of a well containing exactly one cell, given that it contains any cells ($p(k = 1,\lambda \,|\, k > 0)$). (You can arrive at a simpler expression by using $p(k > 0) = 1 - p(k = 0)$).

$$p(k = 1,\lambda \,|\, k > 0) = \frac{p(k=1)}{p(k>0)} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} = \frac{\lambda}{e^{\lambda} - 1}$$

b) What range for λ ensures that this probability is at least 95%?
Expression above > 0.95. $0 < \lambda < 0.102$

c) You manually count the number of cells per well across 1000 wells and want to test whether the numbers are consistent with a Poisson distribution. Describe how you can do this, along with the steps of the strategy you employ.
KS test. Provide some description of the process.

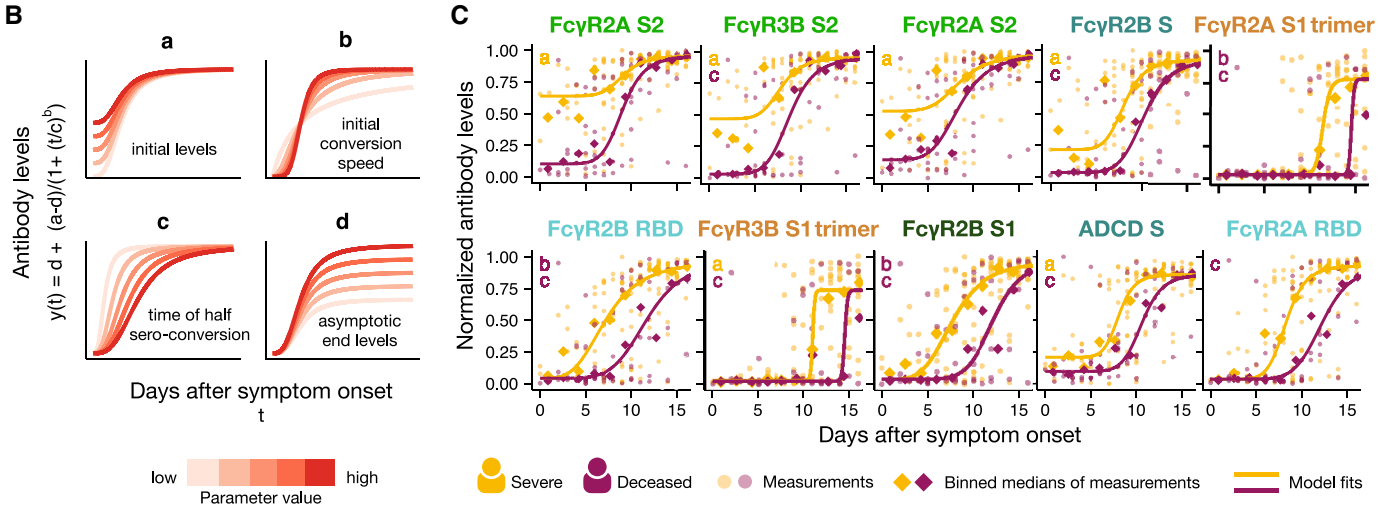


d) What are three things you can say about the distribution of cells per 96-well plate, compared to the distribution of cells per well?
Key is that this is a sample distribution. Tends toward the same mean at 96 times the single well mean. Closer to normally distributed. Smaller variance relative to the mean.

e) See the diagram of the Poisson distribution on the right. Does this distribution have positive or negative skew? Explain.
Positive skew. The distribution tail is heavier on the positive side.

Question 2 (10 points)

Zohar *et al*, *Cell*, 2020 analyzes the dynamics of antibody response during SARS-CoV-2 infection. Each of the plots below depicts a certain measure of antibody quality/quantity, separated by whether subjects survived after being admitted to the ICU. They summarize these dynamics by fitting the longitudinal data to a logistic curve model using non-linear least squares (NNLS).



a) What do you need to provide to NNLS that you do not need to define for ordinary least squares?
A function to describe the input-output relationship. A fitting starting point for the parameters.

b) What are three specific core assumptions of non-linear least squares?
(1) Independent data points. (2) The input-output relationship defined by the provided function. (3) Normally distributed error around the line of prediction.

c) Zohar *et al* aligns samples according to days after symptom onset by asking patients for how many days they have had symptoms on admission. This introduces potential error in the x-position of the measurements. There is also error in each measurement due to biology and technical issues, which shows up in the y-position of each point. Which, if any, of these errors are modeled by the NNLS process? If one of these is modeled, how can you determine the model's estimate of that error?
NNLS only models error in the y direction. The model's estimate of this error is the standard deviation of the residuals.

d) Do you expect some points affect the model solution more so than others? If so, which?
Points far away from the prediction line affect the fit more.

e) You bootstrap your model to examine the uncertainty in the b parameter. How would you expect the b parameter to be distributed, given that you have a large number of data points?
Close to normally distributed.

f) You wish to cross-validate your model to verify that it is not over-fit to the data. In the study, each subject was measured at 2–3 timepoints. Should you split the data on a per-sample or per-subject basis? Explain your reasoning.
You should cross-validate by splitting across subjects, because samples from the same subject are not independent.

Question 3 (10 points)

a) What does bootstrapping provide? How does one perform it?
Estimates model variance. You repeatedly resample your original dataset *with replacement*, so that the dataset is the same size.

b) What does cross-validation aim to estimate? How are its estimates consistently different from the true value? Why is/isn't this concerning?
Aims to estimate the prediction error. Estimate is consistently an overestimate, because the dataset is made slightly smaller. This isn't concerning because it ensures we will tend to see our model work better than estimated.

c) You have a regression model that fits your data well but has poor performance upon cross-validation. Explain what is happening here. What are two classes of approaches you could use to fix this?
The model has high variance, indicating it is perhaps overfit, or the data is insufficient for the model. Two approaches we could use to fix this are regularization and/or dimensionality reduction.

d) You and your research team are excited to have developed a LASSO model that uses mutations from a patient's tumor to predict their response to a drug. The model fits and predicts well on cross-validation. Upon fitting the model, it selects an intriguing mutation in gene X (this gene has a strong coefficient weight). However, upon bootstrapping gene X has a large coefficient about 20% of the time. What do you make of this? Would it be safe to say the model indicates that gene X is essential to predicting drug response?
The model is consistently predictive, but it does not consistently pick this gene to perform these predictions. Therefore, gene X is not *essential* for predicting drug response.

e) Outline the steps for performing cross-validation of a model in order, including normalizing your data by z-scoring.
(1) Split your data into training and test sets. (2) Normalize. (3) Fit. (4) Predict the left out data. (5) Compare the predictions and left out data. (6) Go back to 1 with a new split.

f) Why are multiple folds necessary during cross-validation? What are two trade-offs when selecting a number of folds?
If we only use one fold, the results can be highly dependent on exactly what data we held out. Multiple folds eliminates this effect, and provides us a better prediction of the prediction error overall. More folds can provide us a better estimate of the prediction error (because the effect from part (b) is reduced), but can take much longer to calculate.

Question 4 (10 points)

Ford *et al*, *Clin Infect Dis*, 2021, report that a SARS-CoV-2 rapid test has a sensitivity of roughly 80% in symptomatic people and 40% in asymptomatic people. The specificity was determined to be more than 99.5%.

a) Write out Bayes law, and then rewrite the equations to reflect the probability of an individual actually being SARS-CoV-2 negative, given they had a negative test result.

$$p(A\,|\,B) = \frac{p(B\,|\,A)\,p(A)}{p(B)}$$

A stands for actually. T stands for test.

$$p(A_-\,|\,T_-) = \frac{p(T_-\,|\,A_-)\,p(A_-)}{p(T_-)}$$

b) The incidence of SARS-CoV-2 in Los Angeles on this day overall is 2 in 1000. 10% of those with related symptoms are turning out to be positive for SARS-CoV-2. Calculate the probability of a symptomatic person being negative, given they test negative on a rapid test. Is a tested symptomatic, or untested asymptomatic, individual more likely to be negative?

One estimate of the chance an untested, asymptomatic individual might have SARS-CoV-2 is the overall incidence rate. So, the chance of them *not* having it would be 99.8%. Answers could vary here, because if most cases are symptomatic, that would drive this percentage up.

For a symptomatic individual:

$$p(A_-\,|\,T_-) = \frac{p(T_-\,|\,A_-)\,p(A_-)}{p(T_-)} = \frac{p(T_-\,|\,A_-)\,p(A_-)}{p(T_-\,|\,A_-)\,p(A_-) + p(T_-\,|\,A_+)\,p(A_+)}$$

$$p(T_-\,|\,A_-)\,p(A_-) = (0.995)(0.9) = 0.8955$$

$$p(T_-\,|\,A_+)\,p(A_+) = (0.2)(0.1) = 0.02$$

$$p(A_-\,|\,T_-) = 97.8\,\%$$

A symptomatic individual is more likely to have SARS-CoV-2, even if they test negative.

You are working on deploying a new medical device in hospitals and want to ensure there are sufficient backups in place in case one fails. To understand this, you want to model the amount of time it takes a device to fail. You expect that failures are completely random in time, and so you model the time to failure as an exponential distribution (this models the time to a Poisson-distributed event):

$$p(t) = \lambda e^{-\lambda t}$$

c) You want to use $p(\lambda) = 1/\lambda$ as your prior expectation of the failure rate (in units of years). So far, one device failed at 1 year, and another at 2 years. Derive an expression for the posterior distribution of the failure rate.

First use Bayes Law.

$$p(\lambda \mid t) = \frac{p(t \mid \lambda)\, p(\lambda)}{p(t)} = p(t \mid \lambda)\, p(\lambda)$$

$$p(\lambda \mid t) = \lambda e^{-\lambda} \lambda e^{-2\lambda} \frac{1}{\lambda} = \lambda e^{-3\lambda}$$

$$\int_0^\infty \lambda e^{-3\lambda} \delta\lambda = 1/9$$

$$p(\lambda \mid t) = 9\lambda e^{-3\lambda}$$

d) For a certain failure rate, the following Binomial distribution-based expression gives the probability of 2 devices out of 10 failing within a single year:

$$p(\lambda) = 45\lambda^2(1-\lambda)^8$$

Derive an expression for the chance of seeing two devices fail in a given year, given your observations in (c). (You do not have to solve.)

$$\int p(\text{two devices} \mid \lambda)\, p(\lambda \mid t) \delta\lambda$$

$$\int_0^\infty \left[45\lambda^2(1-\lambda)^8\right] \left[9\lambda e^{-3\lambda}\right] \delta\lambda = 405 \int_0^\infty \lambda^3 (1-\lambda)^8 e^{-3\lambda} \delta\lambda$$

Question 5 (10 points)

You wish to model the dynamics of an outbreak with a susceptible/infectious/recovered (SIR) model.

$$\frac{\delta S}{\delta t} = -\beta SI \qquad \frac{\delta I}{\delta t} = \beta SI - \alpha I \qquad \frac{\delta R}{\delta t} = \alpha I$$

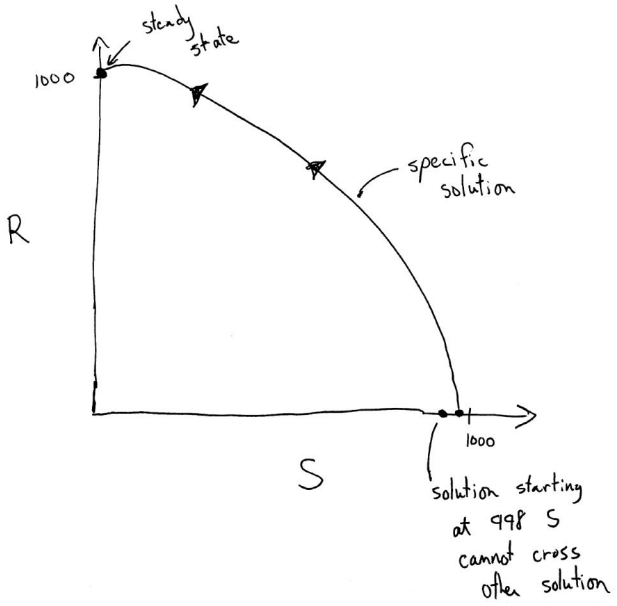
The overall population ($S + I + R$) remains constant with 1000 individuals. Initially there is 1 infectious individual. β describes the infectiousness of the agent, and α describes the rate of recovery.

a) What is the steady-state of this system? What is the Jacobian matrix of the system?
The steady-state is everyone is converted to recovered.

$$J = \begin{bmatrix} -\beta I & -\beta S & 0 \\ \beta I & \beta S - \alpha & 0 \\ 0 & \alpha & 0 \end{bmatrix}$$

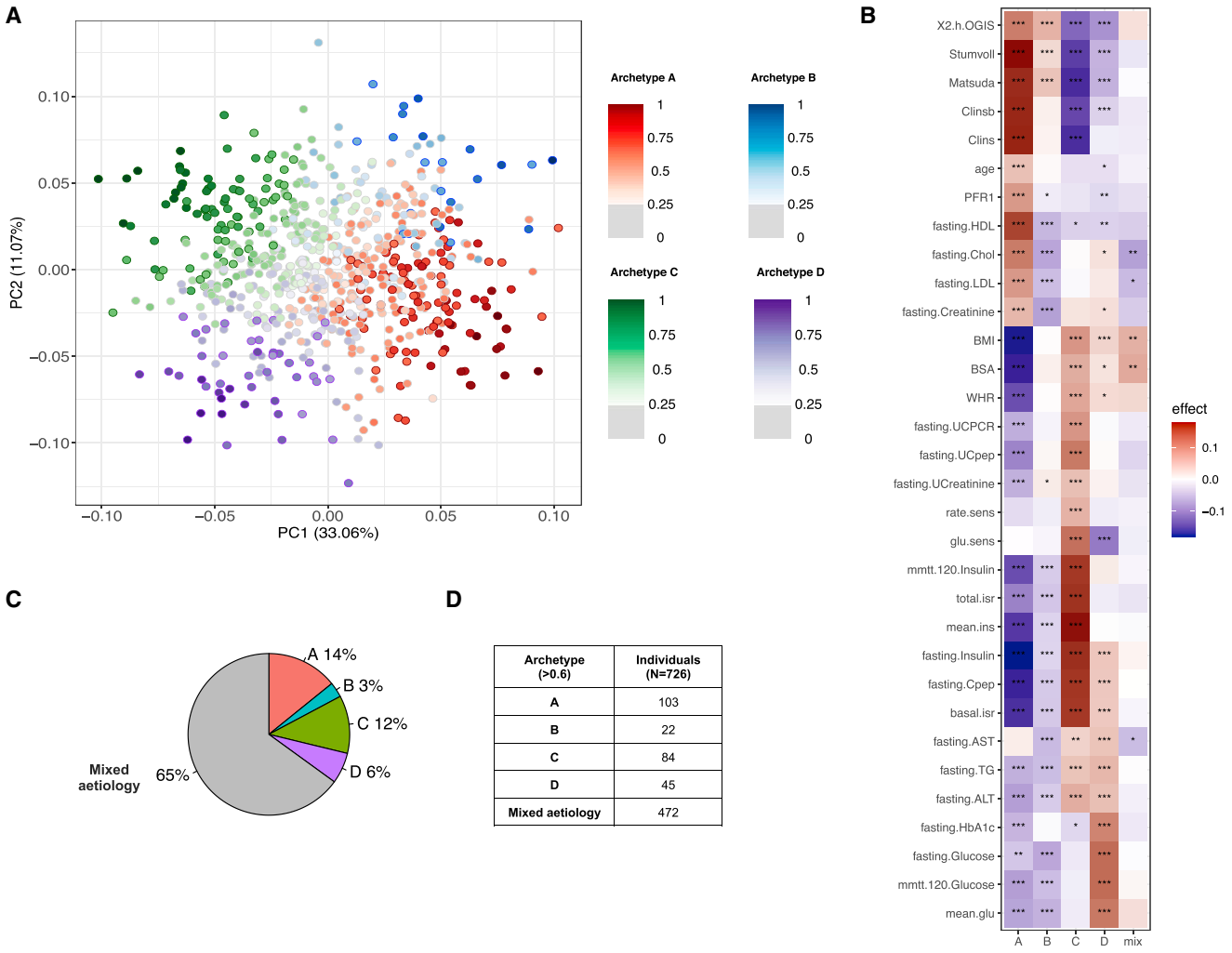
b) Can this system ever lead to oscillatory behavior (spirals or orbits)? If so, for what range of parameters? You may use Wolfram Alpha or another tool to solve for the eigenvalues of the system.
The eigenvalues are real and negative. There is no possibility of oscillatory behavior.

c) Because the number of individuals in the population remains constant, we can draw a state-space diagram with just S and R, ignoring I, and all the properties of this diagram will hold. Sketch such a diagram with the properties you have solved for above. What can you say about the ODE solution starting with 2 infectious individuals, relative to the one with 1 infectious individual?



Question 6 (10 points)

Wesolowska-Andersen *et al*, *Cell Reports Medicine*, 2022 uses a large panel of molecular and clinical characteristics to explore patient-to-patient variation in type II diabetes. Using a variety of analysis techniques, they conclude that four subtypes of the disease exist with differing molecular and clinical patterns. A principal component analysis (PCA) they conducted is reproduced below.



Archetype (>0.6)	Individuals (N=726)
A	103
B	22
C	84
D	45
Mixed aetiology	472

Figure 2. Clinical characteristics of the four archetypes, and groups with archetype scores identified at the extremes of the baseline phenotype spectrum

(A) Representation of the baseline phenotype spectrum of newly diagnosed T2D projected in 2 dimensions following principal-component analysis. Each point represents an individual, and the four archetypes are colored and marked as subgroups A–D. The strength of the colors represents the level of archetype membership, with individuals shown in a lighter color representing a mixed phenotype with no clearly dominating archetype.

(B) Summary of the 32 clinical variables used for the characterization of the baseline T2D phenotypic space. All variables were rank-normally transformed, and for each group with extreme archetype scores and each variable, the heatmap shows the significance level of the difference between the group and the remaining individuals from the study, as from a Mann-Whitney U test. The color of the heatmap reflects the directionality and magnitude of the test estimate, with red indicating higher values and blue indicating lower values characteristic of the given group.

(C) Pie chart showing the percentage of individuals belonging to each of the four groups with extreme archetype scores and in the mixed etiology group.

(D) Table of the number of individuals represented in each of the four groups with extreme archetype scores and in the mixed etiology group.

Values statistically different from zero are marked as *p < 0.05, **p < 0.01, and ***p < 0.001.

a) Given what you know about PCA, what are two model-related things you can say about the patterns represented by PC1 and PC2 here? (Shouldn't be about diabetes.)

PC1 represents the direction of maximum variation. PC1 and PC2 are orthogonal. (Could be other answers.)

b) What is the range of possible variance that might be explained by a third principal component? How would performing PCA with three principal components change the two PCs shown above?
0–11%. A third component wouldn't change the first two.

c) Fasting insulin levels are higher in subtypes C and D. Where would you expect to find this variable on a plot of the loadings for PCs 1 and 2?
Negative along PC1. Could vary with respect to PC2.

d) You re-implement the authors' analysis and find generally similar results, except that archetypes C and D are positive, and archetypes A and B negative, along PC1. Do you have similar results to the authors? What do you expect the loadings to look like relative to the authors'?
These are the same results as the authors'. PCA is sign-indeterminant, and so the results can be flipped about an axis. If this occurs in the scores, then the same reflection should appear in the loadings.

e) You use partial least squares regression instead of PCA to predict the amount of insulin use required by subjects. What can you say about whether/how much the X scores above would change?
If the directions of maximum variance in X happened to be aligned with the directions of covariation, then the results could be the same. In practice, the results could be slightly or greatly different.