Assignment 3.2

Computational Biology (BIOSC 1540)

Due: Nov 7, 2025 by 11:59 pm

3 points

Tier	**C**	**B**	**A**	**S**
Quantity	9	4	3	8
Points	0.24	0.11	0.08	0.02

Problem 1

A microbiologist is growing bacterial colonies on petri dishes. They count the number of colonies on 4 separate dishes prepared from the same source and get the following counts: [6, 4, 7, 5].

This type of count data can often be modeled by a Poisson distribution, which has one parameter, λ, representing the average number of events (colonies). The probability of observing k colonies is given by the formula:

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

where $k!$ is the factorial of k. You are testing two hypotheses from your lab partners:

- **Hypothesis A:** The true average rate of contamination is $\lambda = 5.0$.
- **Hypothesis B:** The true average rate of contamination is $\lambda = 5.3$.

Part A *(C-tier)*

What is the simple average (i.e., the mean) of the observed data? Which hypothesis is this average closer to?

Part B *(C-tier)*

Calculate the total likelihood for each hypothesis. Which hypothesis is the Maximum Likelihood Estimate (MLE) among these two choices? Explain why.

Problem 2

A scientist is modeling the growth of a plant (in cm) as a function of the amount of fertilizer (in mg). They have 10 data points and are trying to fit a regression model. They show you two possible models fit to the *exact same* 10 data points.

- **Model A:** A simple, straight line (linear regression) that captures the general upward trend but does not pass through every point.
- **Model B:** A complex, “wiggly” polynomial line that goes exactly through every data point.

Part A *(C-tier)*

Which model (A or B) will have the lowest Sum of Squared Errors (SSE)? Explain.

Part B *(C-tier)*

The scientist then collects 5 new data points. They find that these new points are very close to the line in Model A, but very far away from the line in Model B. What is the phenomenon that Model B is suffering from called?

Part C *(B-tier)*

Which model (A or B) would you trust to make predictions on future, unseen data? Why?

Problem 3 *(B-tier)*

As a computational biologist, you are given three different datasets from your colleagues. For each scenario, identify the most appropriate probability distribution (Binomial, Poisson, or Gaussian) to model the underlying stochastic process.

1. A scientist uses CRISPR to edit a gene in a dish of 500 cells. The data is a single number: the count of cells that were successfully edited.
2. A geneticist sequences a 1 million base-pair region of DNA from a patient. The data is a single number: the count of new mutations found in that region.
3. A biochemist measures the concentration of a specific protein in 1,000 different samples from a cell culture.

For each of the three scenarios, state which distribution is the best fit and write one sentence explaining why the underlying biological process matches the assumptions of that distribution.

Problem 4

A researcher is studying a rare genetic disease. They believe the disease is caused by a single mutation. They collect a sample of 2,000 people and find that 3 people have this mutation. From this, they calculate the probability of having the disease as $p = \frac{3}{2000} = 0.0015$.

Part A *(C-tier)*

Is $p = 0.0015$ the true probability of having the disease in the entire human population? Explain your answer in the context of a sample vs. the underlying process.

Part B *(A-tier)*

What probability distribution would you use to model the process of observing k individuals with the mutation in a sample of n people?

Part C *(B-tier)*

Imagine the researcher collected a new sample of 2,000 people from the same population. Would you be surprised if they found 4 people with the mutation this time? Or 2 people? Why does this not contradict the first finding?

Problem 5

The Sum of Squared Errors (SSE) is a common loss function, but it's not the only one. Another common choice is the Sum of Absolute Errors (SAE).

SSE = $\sum\left(Y_{\text{pred}} - Y_{\text{obs}}\right)^2$ SAE = $\sum\left|Y_{\text{pred}} - Y_{\text{obs}}\right|$

Let's test two competing models on a small dataset.

X	1	2	3	4
Y_{obs}	1	2	2	10

Model A: $Y = X$ **Model B:** $Y = 2X$

Part A *(C-tier)*

Calculate the SSE for Model A and Model B. According to the SSE, which model is "best"?

Part B *(C-tier)*

Calculate the SAE for Model A and Model B. According to the SAE, which model is “best”?

Part C *(B-tier)*

Did your answer change? What does this tell you about the claim that a model is “the best”? What “hidden assumption” does a scientist make when they choose to use SSE?

Problem 6

The real challenge in RNA-seq is that different transcripts (isoforms) often share parts of their sequences. This creates ambiguous reads. Suppose you are studying a gene that produces two different transcripts, Isoform 1 and Isoform 2.

Isoform 1	**Isoform 2**
[Exon 1]—[Exon 2]—[Exon 3]	[Exon 1]—[Exon 2]—[Exon 4]

Your experiment gives you three types of reads.

Read A	**Read B**	**Read C**
Maps only to Exon 3	Maps only to Exon 4	Maps to Exon 1 and Exon 2

Part A *(C-tier)*

Read A is a unique read. Which isoform did it come from?

Part B *(C-tier)*

Read B is also a unique read. Which isoform did it come from?

Part C *(A-tier)*

Read C is an ambiguous read. Why can't you be 100% certain which isoform it came from?

Part D *(A-tier)*

What is the latent variable for Read C that we wish we knew?

Problem 7

You are trying to quantify the gene expression from a cell with only three transcripts: A, B, and C. You perform a RNA sequencing experiment and obtain read counts for four types of reads, j. You observe each read counts of $N_1 = 30$, $N_2 = 20$, $N_3 = 10$, and $N_4 = 40$.

You then perform a heuristic search of each read and determine which transcript they are compatible with. You store your results in a compatibility matrix, Z_{kj}, which is a binary matrix where $Z_{kj} = 1$ if reads of type j are compatible with transcript k, and 0 otherwise.

Compatibility Z_{kj}	**Read Type 1**	**Read Type 2**	**Read Type 3**	**Read Type 4**
Transcript A	1	0	0	1
Transcript B	0	1	0	1
Transcript C	0	0	1	0

We'll start the EM algorithm with a naive uniform guess for the abundances (fraction) π_k for each transcript k: $\pi_A = 1/3$, $\pi_B = 1/3$, $\pi_C = 1/3$

Part A *(S-tier)*

The E-Step calculates the responsibility, γ_{kj}, which is the probability that a read of type j belongs to transcript k, given our current abundance estimates π_k. The formula is:

$$\gamma_{kj} = \frac{\pi_k Z_{kj}}{\sum_{k'} \pi_{k'} Z_{k'j}}$$

Calculate the responsibility γ_{kj} for all transcripts and all read types. Fill in the table below.

Responsibility γ_{kj}	**Read Type 1**	**Read Type 2**	**Read Type 3**	**Read Type 4**
Transcript A	1.0	0.0		
Transcript B	0.0	1.0		
Transcript C				

Part B *(S-tier)*

The M-Step uses the responsibilities γ_{kj} and read counts N_j to calculate new expected counts $\hat{N}_k$ for each transcript.

$$\hat{N}_k = \sum_j N_j \gamma_{kj}$$

Please fill in the table below and compute the sum across each row.

	$\boldsymbol{N_1\gamma_{k,1}}$	$\boldsymbol{N_2\gamma_{k,2}}$	$\boldsymbol{N_3\gamma_{k,3}}$	$\boldsymbol{N_4\gamma_{k,4}}$	**Sum**
A					
B					
C	0	0	10	0	10

Part C *(S-tier)*

Now, normalize the expected counts to get the new abundance estimates.

$$\pi_k^{\text{new}} = \frac{\hat{N}_k}{\sum \hat{N}_k}$$

π_A^{new}	π_B^{new}	π_C^{new}

Part D *(S-tier)*

Recalculate the responsibility γ_{kj} for all transcripts and all read types with your new responsibilities from Part C.

Responsibility γ_{kj}	**Read Type 1**	**Read Type 2**	**Read Type 3**	**Read Type 4**
Transcript A				
Transcript B				
Transcript C				

Part E *(S-tier)*

Repeat the M-Step with your new responsibilities γ_{kj} and read counts N_j.

	$N_1\gamma_{k,1}$	$N_2\gamma_{k,2}$	$N_3\gamma_{k,3}$	$N_4\gamma_{k,4}$	**Sum**
A					
B					
C					

Part F *(S-tier)*

Re-normalize the expected counts to get the new new abundance estimates.

$$\pi_k^{\text{new}} = \frac{\hat{N}_k}{\sum \hat{N}_k}$$

π_A^{new}	π_B^{new}	π_C^{new}

Part F *(S-tier)*

How would you expect π_C to change if we continued more iterations?

Part G *(S-tier)*

How would you expect π_A and π_B to change if we continued more iterations?