**2.** Aphids are tiny insects that feed on plants such as cabbage plants. A farmer wants to reduce the number of aphids in a cabbage field. A river is located 100 meters south of the cabbage field. The farmer divides the field into 25 regions of equal size, as shown in the diagram. Each region has approximately the same number of cabbage plants.

Farmer's House and Cabbage Field

100 meters
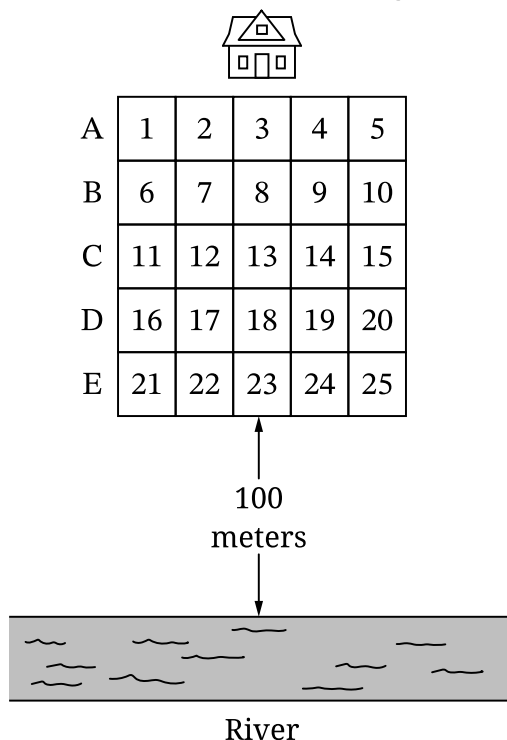
River

The farmer would like to estimate the proportion of cabbage plants in the field that are affected by aphids and believes that the extent of aphid damage is greater for the regions in the cabbage field closer to the river. To obtain the estimate, the farmer is considering three sampling methods.

- Sampling method I: Select region 3, which is closest to the farmer's house and farthest from the river. Examine every cabbage plant in the region for aphid damage.

- Sampling method II: Randomly select one row (A, B, C, D, or E). For every region in the selected row, examine every cabbage plant for aphid damage.

- Sampling method III: Randomly select one region from each of rows A, B, C, D, and E. For each selected region, examine every cabbage plant for aphid damage.

A. Explain whether sampling method I is an appropriate sampling method for the farmer to use to estimate the proportion of cabbage plants in the field that are damaged by aphids.

B. Using sampling method II, the farmer randomly selected row E and examined every cabbage plant in row E. If the farmer's belief is correct, determine whether the selection of row E is likely to provide an overestimate or an underestimate of the proportion of cabbage plants in the field that are damaged by aphids. Justify your answer.

C. Using the information provided in the diagram of the cabbage field, describe how to implement sampling method III, which requires a random selection of one region from each of rows A, B, C, D, and E.

3. Ms. Fey is a manager at a restaurant. To improve the dining experience for her customers, she uses a digital music service to create a playlist of songs that will be played in the restaurant. The playlist contains 1,000 songs and consists of four different types of music in the following quantities: 200 country songs, 400 pop songs, 100 rock songs, and 300 jazz songs. The digital music service will select songs at random from the playlist to be played in the restaurant. Any song can be replayed at any time.

A.

   i. Suppose one song is selected at random to be played. What is the probability that the song is a rock song? Show your work.

  ii. Suppose two songs are selected at random to be played. What is the probability that both songs are rock songs? Show your work.

B. In every one-hour period, 20 songs will be played at random and any song can be replayed at any time. Ms. Fey is interested in how many rock songs will be played in a typical one-hour period.

   i. Define the random variable of interest to Ms. Fey, and state how the random variable is distributed.

  ii. What is the expected value for the random variable in part B (i)? Show your work.

C. Recall that in every one-hour period, 20 songs will be played at random and any song can be replayed at any time.

   i. Determine the probability that 4 or more rock songs in a particular one-hour period will be played. Show your work.

  ii. Suppose 4 rock songs are played during a particular one-hour period. Does this provide strong evidence that the song selection process was not truly random? Justify your answer without performing an inference procedure.

Question 2: Focus on Sampling and Experimental Design — 4 points

General Scoring Notes

- Each part of the question (indicated by a letter) is initially scored by determining if it meets the criteria for essentially correct (E), partially correct (P), or incorrect (I). The response is then categorized based on the scores assigned to each letter part and awarded an integer score between 0 and 4 (see the table at the end of the question).
- The model solution represents an ideal response to each part of the question and the scoring criteria identify the specific components of the model solution that are used to determine the score.

Model Solution	Scoring

A Sampling method I is not an appropriate sampling method for the farmer to use to estimate the proportion of cabbage plants in the field that are affected by aphids.
 	
Sampling method I is a convenience sample where region 3 is not selected randomly. If the farmer's belief is correct, there may be fewer cabbage plants that are affected by aphids in region 3 than in most other regions of the cabbage field because region 3 is in the row farthest from the river. This may lead to an underestimate of the proportion of cabbage plants in the field that are damaged by aphids.
	Essentially correct (E) if the response satisfies the following three components:
	1. Indicates sampling method I is not an appropriate sampling method to obtain the estimate
	2. Provides an explanation that refers to at least one of the following:
	  • The region is likely not representative of the entire cabbage field.
	  • The sampling process does not include random selection.
	  • The sample is a convenience sample.
	3. Provides sufficient context by including any two of the following:
	  • Sampling unit (region)
	  • Population (cabbage field or the 25 regions in the field)
	  • Statistic or parameter (proportion of cabbage plants that are damaged by aphids) OR count (number of cabbage plants that are damaged by aphids)
	
	Partially correct (P) if the response satisfies component 1 AND either component 2 or component 3
	OR
	if the response does not indicate whether or not sampling method I is an appropriate sampling method AND satisfies both components 2 and 3.
	
	Incorrect (I) if the response does not meet the criteria for E or P.

Scoring Notes:

References only to "a field" are sufficient to satisfy the population description in component 3.

Model Solution	Scoring
B The selection of row E is likely to provide an overestimate of the proportion of all cabbage plants in the field that are damaged by aphids. If the farmer's belief that the extent of aphid damage is greater for the regions in the cabbage field closer to the river is correct, then row E, which is the row of regions located closest to the river, is likely to have a greater proportion of cabbage plants damaged by aphids than regions farther from the river.	**Essentially correct (E)** if the response satisfies the following three components: 1. Indicates that the selection of row E is likely to produce an overestimate 2. Provides a justification that is based on the location of row E as the row located closest to the river 3. Links the location to why row E is likely to produce an overestimate (e.g., by referring to the farmer's belief that the extent of aphid damage is greater for the regions in the cabbage field closer to the river) **Partially correct (P)** if the response satisfies only two of the three components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

A response that indicates the selection of row E is likely to provide an underestimate may be scored P if the response provides a justification that ignores the farmer's belief (either by omission or by specific mention) but includes a valid reason for why the regions further from the river may have more aphid damage.

Model Solution	Scoring
C The farmer should write the region numbers from row A, 1 through 5, onto same-size slips of paper, then put the numbers into a hat, mix well, and select one of the numbers. The farmer should repeat this process for the region numbers of each of the other rows (i.e., row B, 6 through 10; row C, 11 through 15; row D, 16 through 20; row E, 21 through 25) and select one number from each row. This process will result in the selection of one region from each row. The farmer will examine every cabbage plant in each of the selected regions for aphid damage to determine the proportion of cabbage plants in the selected regions that are damaged by aphids. *Alternative Solution:* The farmer should use a random number generator to generate one two-digit integer from 01 to 05, one two-digit integer from 06 to 10, one two-digit integer from 11 to 15, one two-digit integer from 16 to 20, and one two-digit integer from 21 to 25. For each integer selected, the farmer should select the corresponding numbered region and examine every cabbage plant in each of the selected regions for aphid damage to determine the proportion of cabbage plants in the selected regions that are damaged by aphids.	**Essentially correct (E)** if the response satisfies the following four components: 1. Describes a random selection process that indicates the groupings used for the strata (by rows) 2. Describes how to correctly implement a random selection process for which the selection of regions within each stratum is equally likely 3. Describes a random selection process for which the selections across strata are independent 4. Describes a random selection process that results in the selection of one region from each stratum **Partially correct (P)** if the response satisfies only two or three of the four components required for E. **Incorrect (I)** if the response does not meet the criteria for E or P.

Scoring Notes:

- A response may satisfy component 1 by referring to the five regions in the rows instead of using row letters.
- For responses that use cards or slips of paper:
  - If the number of slips of paper (number of cards) does not equal 5 for each random selection, then component 1 is not satisfied. Slips of paper (cards) do not need to be specifically identified as equally sized.
  - If the response does not describe a thorough mixing (shuffling) of the slips of paper (cards), then component 2 is not satisfied.
- For responses that use a random number generator or table of random digits:
  - If it is not clear that a random selection process allows the selection of regions within each stratum to be equally likely, then component 2 is not satisfied.
  - If the response does not clearly indicate that a random number is generated from the region numbers within each of the five rows (e.g., only describes the generation of five random numbers from the two-digit integers from 01 to 25), then component 4 is not satisfied.

- For responses that use a fair die:
  - If a five-sided fair die is rolled for each of the five rows and the response clearly indicates the region numbers assigned to the values on the die for each roll, then component 4 is satisfied.
  - If a six-sided die is rolled for each of the five rows and the response clearly indicates which number is excluded (e.g., "if a 6 is rolled, roll again until a non-6 number is achieved") AND the response clearly indicates the region numbers (or columns) assigned to the values on the die for each roll, then component 4 is satisfied.
  - If a 25-sided fair die is rolled for each of the five selections, then component 4 is not satisfied without sufficient further justification because the selection of one region from each stratum is not guaranteed.
- If a response describes two separate random selection processes in detail (e.g., describes how to use a random number generator and slips of paper in a hat), score both descriptions according to the four components and use the lower score.
- If a response indicates that separate samples were taken from each of the strata, then component 3 is satisfied.
- A response that selects columns instead of regions within strata cannot satisfy component 3.