Mock Exam - Midterm 2

Instructions

- This exam was given in Winter 2022 as a 90-minute exam. Your exam will be a 50-minute exam, so it will be proportionally shorter.
- Your exam may have some long-answer proof-style questions (similar to homework questions) even though this exam did not. Long-answer proof-style questions will be graded with partial credit. Multiple choice and short-answer questions will be graded on correctness of the answer alone.
- This exam was themed around a common topic. Your exam may not have a theme.
- For Midterm 2 and Final Part 2, you will be allowed any number of handwritten note sheets.
- You will not be allowed to use a calculator, and you will not need it.
- For Midterm 2 and Final Part 2, all answers can be left unsimplified in terms of permutations, combinations, factorials, exponents, etc.

Background

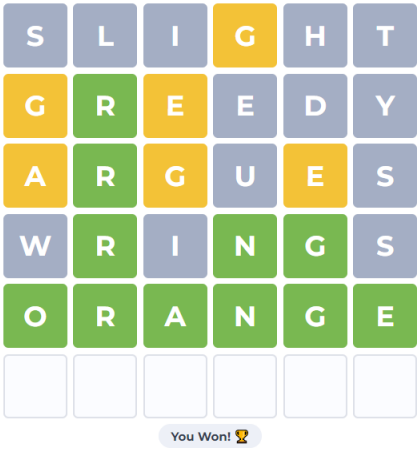
In the game Stringle, players try to guess a randomly generated string. There is a new Stringle string available each day.

Each day's Stringle string is a **six-letter string**, where each letter is chosen uniformly at random, **with replacement**, from among the 26 letters of the English alphabet. This means Stringle strings can have repeated letters, and they do not need to have any meaning as an English word.

Players input a guess, which can be any six-letter string, and get feedback on their guess, which is given by coloring the letters of the guessed string as follows:

- **green** for letters in the correct position,
- **yellow** for letters in the Stringle string, but in the incorrect position,
- **gray** for letters not in the Stringle string.

Based on this feedback, they can guess again, an **unlimited** number of times, until they guess the string correctly, at which point the game is over. Here's an example of what happens when someone plays Stringle and guesses the string on their fifth guess:



A player's Stringle **score** is the number of guesses it took them to guess the word correctly. In the example above, the player's score is 5.

1. (1 point) Stringle uses the following color scheme, or assignment of colors to meanings:

- **green** for letters in the correct position,
- **yellow** for letters in the Stringle string, but in the incorrect position,
- **gray** for letters not in the Stringle string.

Suppose the game developers could have instead used any three colors from the set of seven colors {red, orange, yellow, green, blue, purple, gray}, so long as no color has multiple meanings. How many color schemes include the color purple?

- ○ 15
- ○ 30
- ● 90
- ○ 120
- ○ None of the above.

$C_1^1 \times C_6^2 \times A_3^3$ (15, 6)

2. (1 point) Suppose that every day, the Stringle string is chosen without replacement from among the set of all possible valid strings. What is the probability that DSCTEN was the Stringle string for one of the first 100 days after the game's release?

Solution:

$$\frac{100}{26^6} = 100 * \frac{P(26^6-1,99)}{P(26^6,100)} = 1 - \frac{P(26^6-1,100)}{P(26^6,100)}$$

$26^6 - 1$; $P(26^6, 100)$; $26^6 \times (26^6-1) \times \cdots$ (100 terms); $C(100,1), 26^6$ position.

3. (1 point) What is the probability of today's Stringle string having exactly one duplicate letter, like SYSTEM, BEFORE, or YELLOW?

- ○ $\frac{P(26,5) * 5!}{26^6}$
- ○ $\frac{C(26,5) * 5!}{26^6}$
- ● $\frac{C(6,2) * P(26,5)}{26^6}$
- ○ $\frac{C(26,5)}{26^6}$
- ○ None of the above.

$(26-1) \times (26-2) \times \cdots$ (99 terms)

① 5 unique, $\binom{26}{5}$, $C(26,5)$

② $P(5,5)$ $P(26,5)$

③ $C(6,2)$, picking 2 locations out of 6 letters. $C(6,2)\,C(4,4)$; $\frac{6\times5}{2\times1}$

4. (1 point) What is the probability of today's Stringle string having at least one duplicate letter, like SYSTEM, GOOGLE, or WAHOOO?

Solution:

$$1 - \frac{P(26,6)}{26^6}$$

M I S P: $\binom{11}{1}\binom{10}{4}\binom{6}{4}\binom{2}{2}$

$\frac{11!}{4!4!2!}$ = MISSISSIPPI (M:1, S:4, I:4, P:2)

$\frac{6\times5}{2\times1}$; SSYTEM, SSYTEM, SYTEM; ① ②

$\binom{5}{1}\cdot\binom{6}{1} / P_2^2$; S S; $\frac{\binom{26}{4}}{\binom{2}{2}}$

5. (1 point) What is the probability that today's String string has the last three letters matching the first three letters, like BAMBAM or GOOGOO, if we know that the first letter matches the fourth letter?

- (●) $\frac{1}{26^2}$
- ( ) $\frac{1}{26^3}$
- ( ) $\frac{P(26,3)}{P(26,5)}$
- ( ) $\frac{P(26,3)}{26^6}$
- ( ) None of the above.

$\frac{26 \cdot (26 \cdot 26)}{26^5}$

6. (1 point) How many possible Stringle strings contain exactly three H's, like HAHAHA or SSSHHH?

- ( ) $C(6,3) * 26^3$
- ( ) $C(6,3) * 26^6$
- ( ) $C(6,3)$
- ( ) 25^3
- (●) None of the above.

$C_6^3 \cdot 25 \cdot 25 \cdot 25$

7. (1 point) For this problem, we'll say that there are six vowels: A, E, I, O, U, and Y. Consider the following three events:

- A is the event that today's Stringle string starts with a vowel.
- B is the event that today's Stringle string starts with a letter in the first half of the alphabet (A through M, inclusive).
- C is the event that today's Stringle string does not start with a Z.

Which of the following is true?

- ( ) A and B are independent. A and B are conditionally independent given C.
- (●) A and B are independent. A and B are conditionally dependent given C.
- ( ) A and B are dependent. A and B are conditionally independent given C.
- ( ) A and B are dependent. A and B are conditionally dependent given C.
- ( ) None of the above.

8. (1 point) For this problem, we'll say that there are six vowels: A, E, I, O, U, and Y. What is the probability that today's Stringle string satisfies at least one of the following conditions?

- starts with a vowel
- ends with a vowel
- first two letters are the same

$1 - \frac{20 \cdot 25 \cdot 26 \cdot 26 \cdot 26 \cdot 20}{26^6}$

Solution:
$$\frac{6}{26} + \frac{6}{26} + \frac{1}{26} - \frac{6}{26} * \frac{6}{26} - \frac{6}{26} * \frac{1}{26} - \frac{6}{26} * \frac{1}{26} + \frac{6}{26} * \frac{6}{26} * \frac{1}{26} = 1 - \frac{20}{26} * \frac{20}{26} * \frac{25}{26}$$

9. (1 point) Your first guess for today's Stringle has six distinct letters. There are 3^6 possible colored feedback sequences for this first guess, since each of the six letters could be colored one of three colors: green, yellow, or gray. How many of the 3^6 possible feedback sequences for your first guess include 2 green letters, 2 yellow letters, and 2 gray letters?

◯ 6!

◯ $C(6.3)$

◯ $(C(6,2))^3$

● $C(6,2) * C(4,2)$

◯ None of the above.

10. (1 point) Consider the following three events:

- A is the event that today's Stringle string has a duplicate letter.
- B is the event that today's Stringle string starts with a B.
- C is the event that today's Stringle string has the same first and last letter.

Which of the following is true?

◯ A and B are independent.
A and C are independent.
B and C are dependent.

◯ A and B are dependent.
A and C are independent.
B and C are independent.

◯ A and B are independent.
A and C are dependent.
B and C are dependent.

◯ A and B are dependent.
A and C are dependent.
B and C are independent.

● None of the above.

11. (1 point) Suppose your Stringle strategy is to guess a random six-letter string every time. That is, each guess you make is equally likely to be any six-letter string, and you don't incorporate the feedback from your prior guesses into subsequent guesses. You might even guess the same string twice!

With this (horrible) strategy, what is the probability that your Stringle score for today is 8?

Solution:

$$\left(1 - \frac{1}{26^6}\right)^7 * \left(\frac{1}{26^6}\right)$$

12. (1 point) Your competitive family keeps track of their Stringle scores in a spreadsheet, where each column represents a different date, and each person in a the family records their scores in a different row. Suppose your family has 5 people, and you have kept track of Stringle scores for 50 days, so your spreadsheet has 5 rows and 50 columns. How many ways are there to organize the spreadsheet, if the rows can be in any order, but the columns must be in chronological order or reverse chronological order?

Solution:

$$5! * 2 = 240$$

13. (1 point) Suppose you apply k-means clustering to the data in your family's Stringle score spreadsheet by representing the entries in each column of the spreadsheet as a vector in $\mathbb{R}^5$. If we run the algorithm with $k = 2$, which of the following is the best description of what the clusters represent?

● One cluster represents the Stringle strings that were easier to guess, and the other represents the Stringle strings that were harder to guess.

◯ One cluster represents the family members who are better Stringle players, and the other represents the family members who not as good.

◯ One cluster represents the Stringle strings that had duplicated letters, and the other represents the Stringle strings that had distinct letters.

◯ One cluster represents the family members who play Stringle more often, and the other represents the family members who do not play as often.

14. (1 point) You survey 250 Stringle players and ask them several questions about their Stringle-playing behavior, with the goal of figuring out how much they would be willing to pay if Stringle were to switch to a paid annual subscription. All of the questions are yes or no questions, except for

"What's the maximum amount you would pay for a year's subscription to Stringle?"

which has options \$0, \$10, or \$20. Of the 250 participants,

- 100 said they would pay \$0,
- 50 said they would pay \$10, and
- 100 said they would pay \$20.

The table below shows the survey questions and gives the number of people who answered yes to each question, broken down by how much they would pay for a subscription.

Question	Number of yes responses from those who would pay \$0 **(among 100 people)**	Number of yes responses from those who would pay \$10 **(among 50 people)**	Number of yes responses from those who would pay \$20 **(among 100 people)**
Do you play Stringle every day?	30 30%	10 20%	50 50%
Do you share your Stringle results with friends or family?	40 40%	20 40%	40 40% X
Do you record your Stringle scores anywhere?	30 30%	30 60%	30 30%
Have you ever recommended Stringle to someone you know?	40 40%	40 80%	80 80%

Assume that each person answered each question on the survey. For example, of the 100 people who said they would pay \$0 for Stringle, 30 said "Yes" to the question *Do you play Stringle every day?*, and the other 70 said "No".

You will use the results of this survey to try to predict how much a person is willing to pay based on their answers to the four questions listed in the table above, using a naive Bayes classifier (without smoothing).

What would your classifier predict for someone who answers the survey questions as follows:

- *Do you play Stringle every day?* "No"
- *Do you share your Stringle results with friends or family?* "Yes"
- *Do you record your Stringle scores anywhere?* "No"
- *Have you ever recommended Stringle to someone you know?* "No"

Hint: It's not necessary to do lots of arithmetic.

● This person would pay \$0 for a Stringle subscription. ✓

◯ This person would pay \$10 for a Stringle subscription.

◯ This person would pay \$20 for a Stringle subscription.

15. (1 point) This problem is a continuation of the previous problem, using the same survey results.

After making predictions for various individuals based on their answers to all four survey questions, you realize that in every case, regardless of how the person answered the survey questions, you could make the same prediction using only three of the four survey questions. For this data, which survey question can you safely remove, without changing the results of your naive Bayes classifier?

◯ Do you play Stringle every day?

● Do you share your Stringle results with friends or family?

◯ Do you record your Stringle scores anywhere?

◯ Have you ever recommended Stringle to someone you know?

◯ None of the above. Removing any question will change the results of the naive Bayes classifier.

16. (1 point) Stringle has become such a popular game that many spin-offs have been created. In a variant called Quadringle, players have to solve four different Stringle puzzles at the same time. We'll say that a Quadringle puzzle is determined by which four six-letter strings it includes, and the order of those four strings does not matter. How many Quadringle puzzles are possible?

◯ $(26^6)^4$

● $C(26^6, 4)$

◯ $P(26^6, 4)$

◯ $4^{C(26,6)}$

◯ None of the above.

17. (1 point) In today's Quadringle puzzle, the four strings are all very similar:

DEALER, RELAYS, ALLOYS, and SALADS.

Suppose you pick one of these four strings at random, so that each has a $\frac{1}{4}$ probability of being chosen. Then you choose one letter of that word at random, such that each has a $\frac{1}{6}$ probability of being chosen.

It turns out that you picked an L. What is the probability that your L came from the word SALADS?

- ○ $\frac{1}{24}$
- ○ $\frac{1}{6}$
- ● $\frac{1}{5}$
- ○ $\frac{1}{4}$
- ○ None of the above.

$$P(\text{salad} \mid L) = \frac{P(L \mid \text{salad}) \cdot P(\text{salad})}{\Sigma}$$

$$= \frac{\frac{1}{6} \cdot \frac{1}{4}}{\frac{1}{6} \cdot \frac{1}{4} \times 3 + \frac{2}{6} \cdot \frac{1}{4}} = \frac{1}{3+2}$$

18. (1 point) In today's Quadringle puzzle, the four strings are all very similar:

DEALER, RELAYS, ALLOYS, and SALADS.

Suppose you pick one of these four strings at random, so that each has a $\frac{1}{4}$ probability of being chosen. Then you choose one letter of that word at random, such that each **distinct** letter in the word has an equal probability of being chosen.

It turns out that you picked an E. What is the probability that your E came from the word DEALER?

- ○ $\frac{1}{24}$
- ○ $\frac{1}{6}$
- ○ $\frac{1}{5}$
- ○ $\frac{1}{4}$
- ● None of the above.

$$P(\text{Dealer} \mid E) = \frac{P(E \mid \text{Dealer})\, P(\text{Dealer})}{\Sigma}$$

$$= \frac{\frac{1}{5} \cdot \frac{1}{4}}{\frac{1}{5} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4} + 0 + 0}$$

$$=$$

① Midterm 2, Wednesday lec time

② 4

- ① Permutation & Combinatorics
- ② Conditional
- ③ Bayes' Theorem and Law of total probability
- ④ Freq & Bayesian

Coin: Head, Tail

Outcomes: H T H T

↳ 1 0 1 0

p: prob heads P 1-p P 1-p

Freq: 50% ⟶ Law of large number

Bayesian: ① prior distr on p

non-informative prior

② posterior ∝ prior · likelihood

posterior: distr on para; prior: distr on para; likelihood: sample

② likelihood: P(Sample | parameter)

p = 0: likelihood: 0

p = 1/10: likelihood: $\frac{1}{10} \times \frac{9}{10} \times \frac{1}{10} \times \frac{9}{10} = \frac{81}{10^4}$ <<

p = 1/2: likelihood: $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$ >>

* If outcomes aligns with assumptions, higher likelihood.