wordle scoring

introduction

two people play a game where they must guess a five letter word in a maximum of 6 guesses, where every wrong guess returns some hints as to what the right word is. each player uses their own strategy to play, and they each believe their own strategy to be superior. can we help them find out which is the better strategy?

the scores

when the players have played 100 games, and their scores are tallied up (number of times they discovered the true word in k guesses for k in [1..6]).

for **player a,** their game summary looks like:

```
|-------------|-----------|
| num guesses | num games |
|-------------|-----------|
|           1 |         0 |
|           2 |         3 |
|           3 |        31 |
|           4 |        49 |
|           5 |        15 |
|           6 |         2 |
|-------------|-----------|
```

but for **player b** it is:

```
|-------------|-----------|
| num guesses | num games |
|-------------|-----------|
|           1 |         0 |
|           2 |         2 |
|           3 |        37 |
|           4 |        36 |
|           5 |        22 |
|           6 |         3 |
|-------------|-----------|
```

if instead we sort and list the number of guesses each player required to find the word for all 100 games, then for **player a** we'd get:

```
2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4,
4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6
```

and for **player b**:

```
2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4,
4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5,
5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6
```

the question we would like to answer is *which strategy performs better?*

method one: the median number of guesses

definition: given a list of scores, the median score_ is the score that splits the list in even halves, such that half the scores are less than or equal to the median score, and half the scores are greater than or equal to the median score.

hypothesis: the player with the lower median number of guesses has the better strategy.

let us analyse each player's history. for **player a**:

- 34 out of 100 games required 3 guesses or fewer
- 80 out of 100 games required 4 guesses or fewer
- the median number of guesses is 4

and for **player b**:

- 39 out of 100 games required 3 guesses or fewer
- 75 out of 100 games required 4 guesses or fewer
- the median number of guesses equals 4.

conclusion: *a draw*. in half the games, *both players got the right answer in 4 or fewer guesses*, and in the other half of the games both players required 4 or more guesses to find the answer.

method two: the balance point of guesses

definition: the balance point of a list of outcomes is a hypothetical outcome that splits the list evenly, like a fulcrum balancing scales. *hypothesis*: the player with the lower balance point of number of guesses has the better strategy.

though the median number of guesses is 4 for both players, the number of times they achieved that score is not the same between them.

- for **player a**:
    - lower half of the scores: [0, 3, 31, 16, 0, 0]
    - upper half of the scores: [0, 0, 0, 33, 15, 2]
    - the balance point score is 3 + 16/49 = **3.327**
- for **player b**:
    - lower half of the scores: [0, 2, 37, 11, 0, 0]
    - upper half of the scores: [0, 0, 0, 25, 22, 3]
    - the balanced number of guesses is 3 + 11/36 = **3.306**

conclusion: 3.306 < 3.327 so **player b** has a (slightly) lower balance point of the number of guesses. **player b** wins.

method three: average number of guesses

given a list of scores, the median score is generally a good representative statistic because it is not sensitive to rare, anomalous, extreme values. however, in the case at hand, the outcomes are bounded and all in the set of {1,2,3,4,5, and 6}. therefore the arithmetic mean (simple sum of scores divided by count of games) is not a poor choice for a summary statistic.

hypothesis: the player with the lower arithmetic mean number of guesses has the better strategy.

for **player a** the arithmetic mean of the number of guesses is:

```
2+2+2+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3
3+3+3+3+3+3+3+3+3+3+3+3+3+3+4+4+4+4+4+4
4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4
4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4
4+4+4+5+5+5+5+5+5+5+5+5+5+5+5+5+5+5+6+6 / 100
= 3.82
```

meaning **player a**'s strategy yielded the correct answer in, on average, 3.82 guesses.

likewise for **player b**:

```
2+2+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3
3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+3+4
4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+4
4+4+4+4+4+4+4+4+4+4+4+4+4+4+4+5+5+5+5+5
5+5+5+5+5+5+5+5+5+5+5+5+5+5+5+5+5+6+6+6 / 100
= 3.87
```

meaning **player b**'s strategy yielded the correct answer in, on average, 3.87 guesses.

conclusion: **player a** has a very slightly *lower arithmetic mean number of guesses*, and thus a better strategy.

method four: weighted average score

definition: the arithmetic mean is just and blind, as it considers all game outcomes equally important.

on the other hand, a *weighted average* assigns unequal importance to the outcomes, scales the outcomes by their importance, and calculates the net average of the scaled values. the value of this method is that we can encapsulate more information and intuitions about the system we are modeling, but the drawbacks are that the weights could be assigned in many different ways, and thus the choice needs to be carefully justified, and the score is harder to interpret.

this may be a reasonable approach given the fact that the players accumulate more (chances of acquiring) information about the true word as their number of guesses rises. given that the players have more information, a wrong 5th guess is making a larger error than a wrong second guess. we can translate that intuition into a weighted average. looking at it the other way, guessing the right word after two failures is a greater stroke of brilliance *given the information available at the time* than is guessing the right word after 3 or more failures.

there are many ways one could distribute the weights across the scores. one simple way would be to consider every wrong guess as providing the same amount (1 unit) of "information" about the true word, and weigh the scores by the amount of information accumulated at time of guess.

```
|--------------|----------------------|----------------|
| guess number | information available | weighted score |
|--------------|----------------------|----------------|
|            1 |                    0 |              0 |
|            2 |                    1 |              2 |
|            3 |                    2 |              6 |
|            4 |                    3 |             12 |
|            5 |                    4 |             20 |
|            6 |                    5 |             30 |
|--------------|----------------------|----------------|
```

as can be seen by the weighted score above, this method exaggerates the penalty of requiring more guesses (the penalty grows with k as k squared.) a more moderate approach can easily be devised, but is not considered here.

given this (admittedly harsh) weighting, the scores for **player a** get transformed to:

```
 2,  2,  2,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,
 6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6, 12, 12, 12, 12, 12, 12,
12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 30, 30
```

an average weighted penalty score of 11.4

likewise, **player b**'s weighted penalty scores:

```
 2,  2,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,
 6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6,  6, 12,
12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,
12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 20, 20, 20, 20, 20,
20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 30, 30, 30
```

yield an average weighted penalty score of 11.8.

conclusion: 11.4 < 11.8 and **player a**'s strategy makes better use of the available information.

statistical analysis of difference

it is obvious that, in the case at hand, the two players perform at a similar level. though it is useful to rigorously define rule that can determine which is the winner, it is also worth asking whether there is any significant evidence based on past performance, as to whether the two stategies can be confidently said to differ at all. we can probe that question in a couple of different ways:

student's t-test for comparin two sample means

assumption: the student t-test applies when we assume the two samples (sets of scores of **a** and **b**) are each drawn from a normally distributed populations with an unknown mean, and variance. (each draw independent from the previous).

null hypothesis: there is no difference between the average number of guesses required by each of the two players.

alternate hypothesis: either **player a** has higher expected average number of guesses or **player b** has higher expected average number of guesses.

```
assuming equal variance
`t-statistic: -0.4203, p-value=0.67`
relaxing assumption of equal variance doesn't change outcomes:
`t-statistic: -0.4203, pvalue=0.67)`
```

this result tells us that there is not enough evidence in the data to reject the null-hypothesis. i.e. had the two samples been random draws from a normal distribution with the same mean number or guesses, we would expect the data to thus collected generate the same amount of difference between the sets.

kolmogorov-smirnov test of shared population distribution

assumption: we need not assume that the two sets of scores were drawn from a normally distributed population.

null hypothesis: the two sets of scores are drawn from the same population distribution

conclusions

there is no evidence in the available data suggesting statistically significant difference between the performance of the two players. whatever tiny difference is observed, is most likely a result of statistical fluke more than a systematic.

that conclusion stands for the purposes of predicting future performance, but will be unsatifying for the purposes of determining a winner in the game. for determining the winner, we can devise various methods, and associated metrics to compare the players and determine the winner. we have shown the outcome of four such methods above. unfortunately, as the performances are quite close, the metrics of these methods do not agree on the winner.

```
|------------------------|---------------|
|                 metric |        winner |
|------------------------|---------------|
|                 median |        _draw_ |
|          balance point |  **player b** |
|          average score |  **player a** |
| weighted average score |  **player a** |
|------------------------|---------------|
```

which player's strategy performs better depends on the metric we choose. the median cannot distinguish between the two players, but the balance point favours player b. conversely, the average number of guesses required is slightly lower for player a, and that lead increases if we apply a simple but harsh *available_information-penalty-weighting* on the score.

on balance, i declare the narrowest of victories to **player a**.