PART A: Introduction

1/ About the data

In this article, I'll take the national English test in Vietnam as an example. It takes place annually in July and there is only one test for all test takers. The test includes 50 questions shuffled in 24 different ways which then result in 24 exam codes. Each question has 4 answers A, B, C, D and only one key is correct. We're going to work with data in 3 years: 2018, 2019 and 2021.

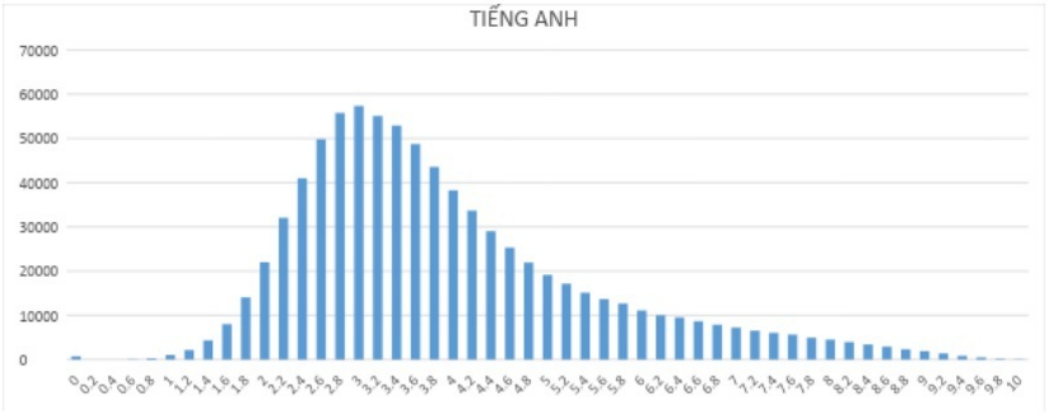
2/ Data collection

The official answer is issued as an image. Unofficial sites (mostly schools, newspapers) then re-post the answer as raw text. I thought of 2 ways to collect the data. One is taking the image from the official site and implementing text recognition. However, the accuracy of text recognition is less than 50%. Using text recognition and then checking each answer would be a tedious task.

The other way is to simply copy and paste the data. I copied all the answers from unofficial sites and prepared them for analysis. The data in 2020 is messy so I didn't use it.

3/ Assumptions and expectations

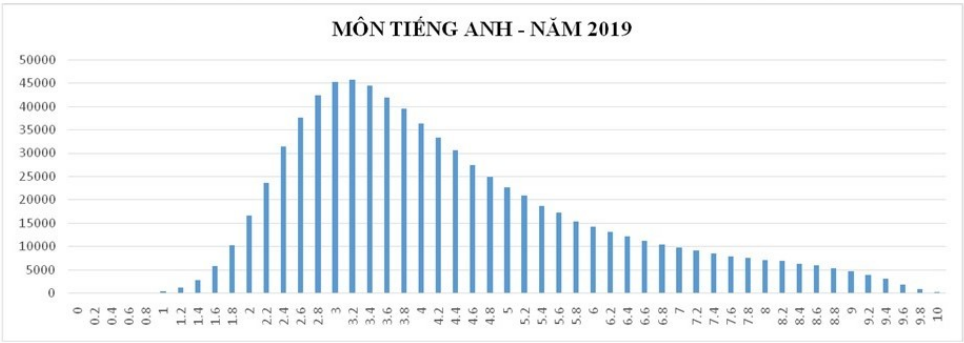
It's impossible to know exactly or to predict the answer for next year. We assume that we are able to get a certain number of correct answers and statistically guess the rest based on "patterns" from previous ones. **In the testing part, I'll take 15 (average) as the certain number because the distributions in 2018, 2019, 2021 say that most students got 15 correct answers.**



2018

9. Tiếng Anh

Điểm	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2
Số lượng	0	0	5	33	123	469	1324	2864	5952	10310	16722	23685	31481	37599	42348	45297	45755
Điểm	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6
Số lượng	44476	41861	39542	36385	33410	30588	27458	24979	22630	20989	18710	17283	15464	14288	13145	12173	11343
Điểm	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8	10
Số lượng	10405	9834	9274	8552	7990	7612	7108	6970	6416	6045	5378	4845	3968	3133	1976	939	299

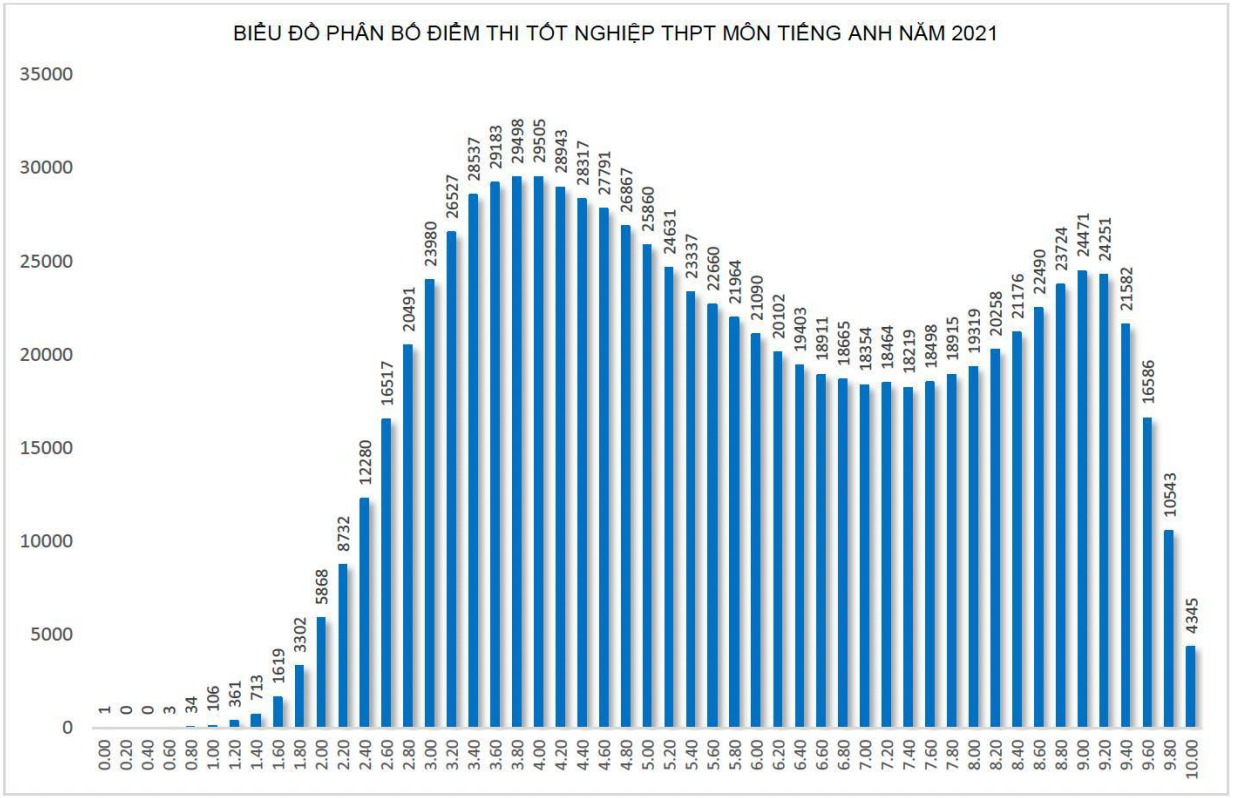


Tổng số thí sinh	789435
Điểm trung bình	4.36
Điểm trung vị	4.00
Số thí sinh đạt <=1 điểm	630
Số thí sinh đạt điểm dưới trung bình (<5 điểm)	542666 (68.74%)
Điểm số có nhiều thí sinh đạt nhất	3.20

2019

9. MÔN TIẾNG ANH

a. Phổ điểm



2021 (it's weird, isn't it?)

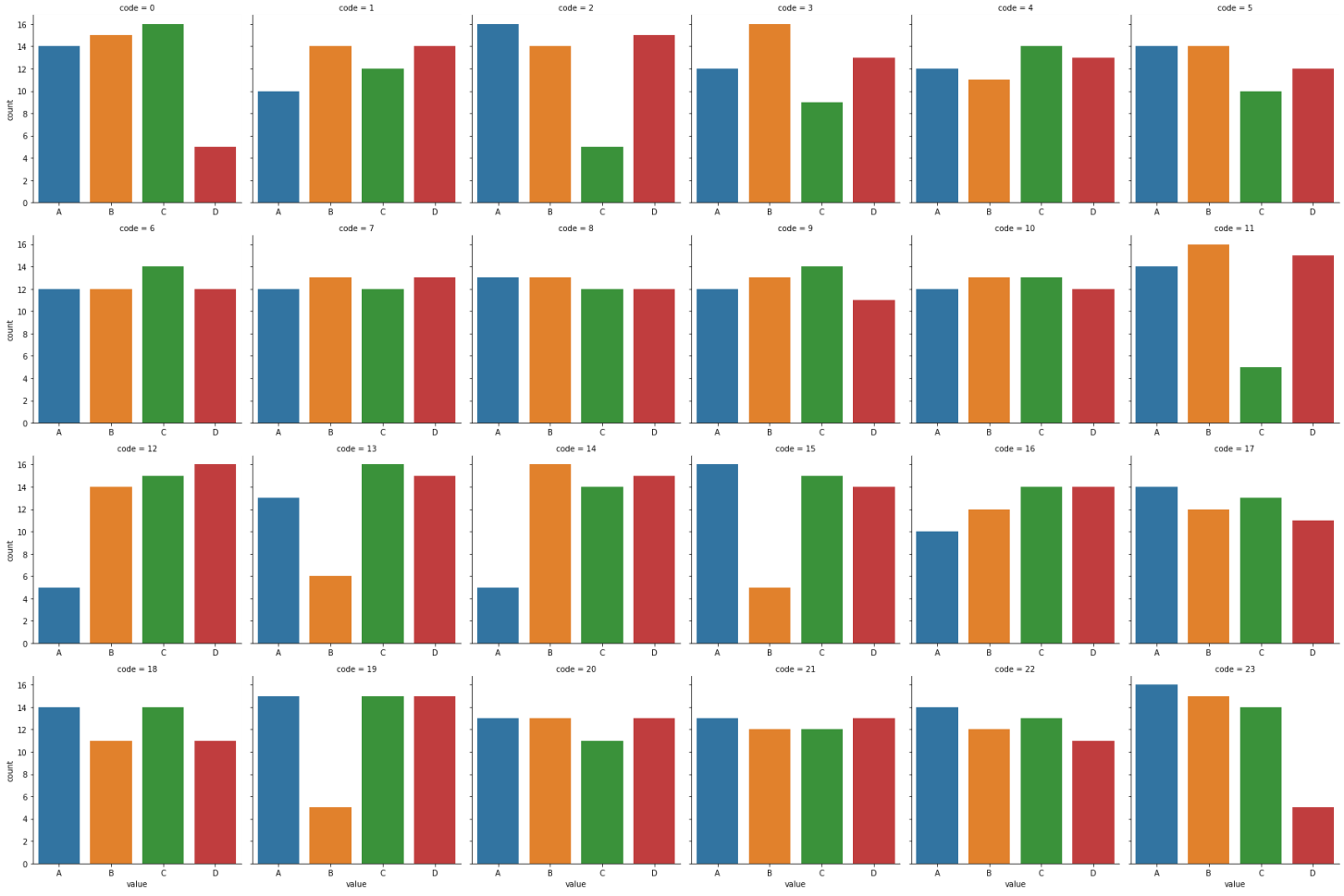
If you know nothing about a question, the probability of getting that one correctly is 0.25. We will try to improve this number. Don't expect a 1.

PART B: Getting started

In this part, we're going to count the number of As, Bs, Cs, Ds and check for the continuity of each key answer in one year 2018. Then, do the same thing in both 3 years to make conclusions.

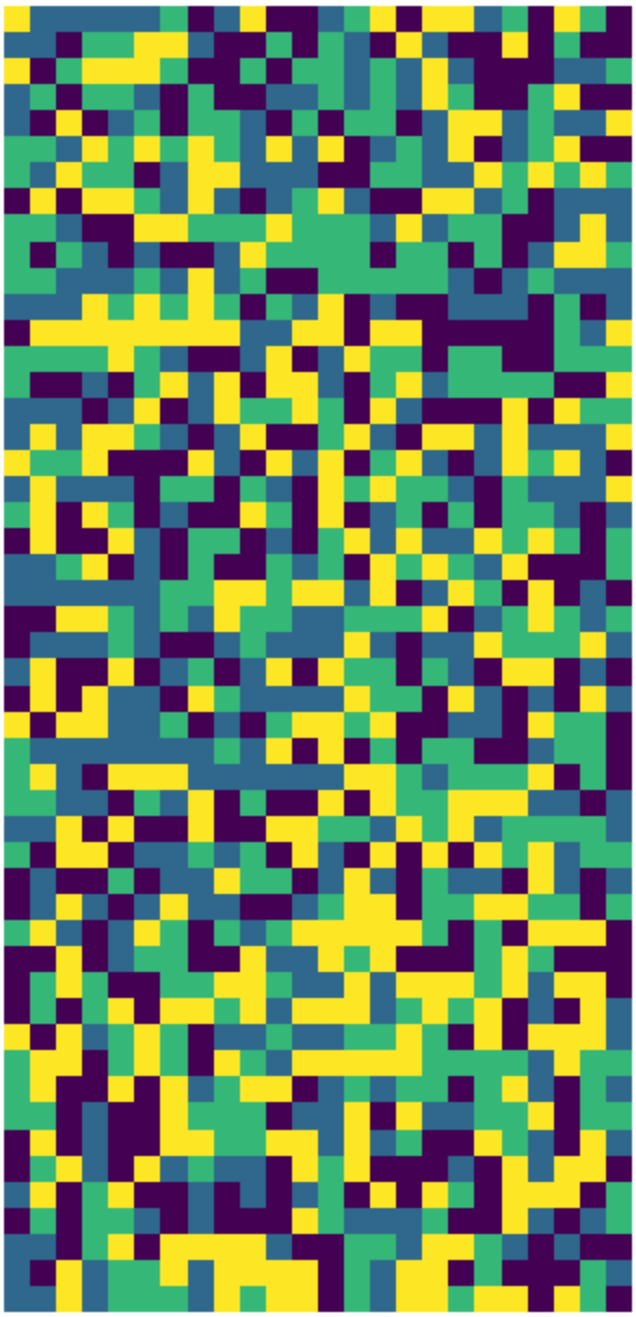
1/ How many As, Bs, Cs and Ds in each exam code? (in 2018)

There are 301, 297, 302, 300 As, Bs, Cs, Ds respectively and below is how it is distributed in 24 exam codes:

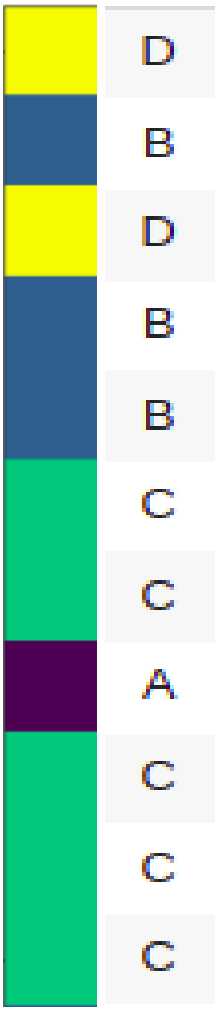


Notice that the number of each letter varies from 5 to 16. This means, for example, you shaded 17 As in your answer sheet, there is a high chance that at least one is incorrect and if you chose only 5 Bs and felt it was so few, it could still be correct.

2/ Next, I'd like to mention continuity. The below image looks like barcodes that represent the answers' continuity in 2018. I'll explain right away. One image has a shape of (50, 24) (50 questions, 24 exam codes), each pixel is one letter and we don't care which colour A,B,C,D is, we only care about their continuity. For instance, looking at the first pixels in the first column (respectively to the first questions of the first exam code), you will see there are three green boxes sticking together which means there are 3 letters coming in a row in the real answer.

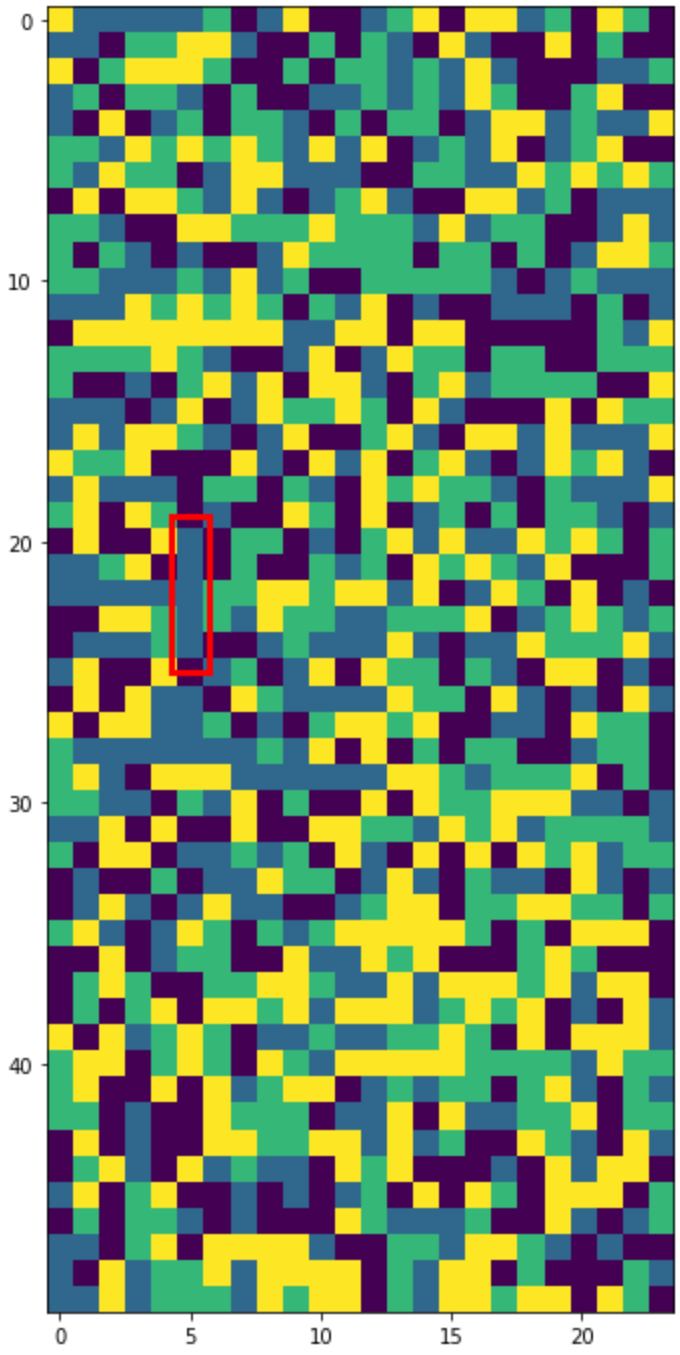


The whole key answers



I zoomed in on the first 11 pixels and compare to the actual answer.

2 or 3 same letters come in a row is natural but how about 4 or 5? You can quickly notice there's no 4 same answers coming in a row, but 5. Frequency? Once.



This is actually a 5 continuous B.

3/ How about 2019 and 2021?

Doing the same thing with 2 other years, we get some information:
(If you're curious about its charts, please navigate to images file)

- The total number of letters is almost equally distributed in one year.
- The total number of each individual letter varies from 5 to 16.
- 4 or 5 same letters coming in a row is very rare.

The total number of As, Bs, Cs and Ds:

		Letters			
		A	B	C	D
Years	2018	301	297	302	300
	2019	305	298	299	298
	2021	301	299	299	301

PART C: Apply our statistics.

Let's say in 2021, Bob is a test taker who is given exam code 0 (the first one). He has finished the first 15 questions and he knows for sure that they are 100% correct. Bob looks at the remaining 35 questions, he absolutely has no idea about those and he can't even eliminate 1 or 2 answers. Bob thinks of 2 strategies: the first strategy is to randomly sparsely shade the answer, the second one is to choose one letter for all and he has 4 sub-options for that.

If he closes his eyes and randomly shades the answer boxes, he might get 0.25 * 35 = 8.75 ~ 8-9 correct answers (The probability of getting one correct is 0.25 and the number of correct answers in the resting questions has the Binomial distribution).

Bob isn't satisfied, he opens his eyes and counts for the letters, he realised so far he has had 7 As, 4 Bs, 2 Cs and 2 Ds. He knows the range of the number of each letter is [5,16] and most of the time, the number of letters are virtually equally distributed so he decides to choose D simply because there are still a lot of Ds. If so, he will get 11 correct answers!! However, is it too early to break out the champagne? What if he all chooses A, B or C for the rest? Here's the outcome:

	Sure	Rest	Expected	A	B	C	D
Number of certain answers	15.0	0.0	0.00	7.00	4.00	2.00	2.00
Number of the rest	0.0	35.0	8.75	4.00	11.00	8.00	11.00
Probability	0.0	0.0	0.25	0.11	0.31	0.23	0.31

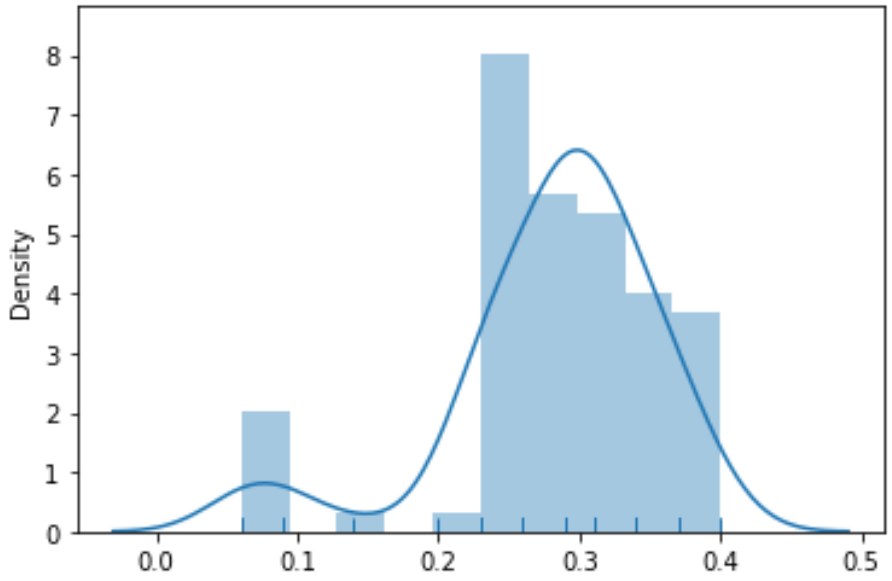
It will be a bit worse if he chooses C. But he has 50% of getting more than 8-9 correct answers.
At the same time in a different room, Bob's friend, Mary is also taking this test. She has 15 questions done but is given code 1. Mary has chosen 4 As, 3 Bs, 4 Cs and 4 Ds. She decides to choose B for all the rest. And

	Sure	Rest	Expected	A	B	C	D
Number of certain answers	15.0	0.0	0.00	4.00	3.00	4.00	4.00
Number of the rest	0.0	35.0	8.75	12.00	11.00	1.00	10.00
Probability	0.0	0.0	0.25	0.34	0.31	0.03	0.29

Great! She's got 11 correctly.

Maybe you want to see all the possible probabilities if you use this method for all exam codes in 3 years. Here I calculated and visualised them for you.

[0.06, 0.09, 0.14, 0.2 , 0.23, 0.26, 0.29, 0.31, 0.34, 0.37, 0.4]



About 66.6% - 77.3% of the time, you will luckily get more than 8-9 correct answers. The average number is 28% equivalent to 9-10 answers. (66% is from normalised data, 77.3% is computed from raw data).

Exception: If you know more than 30 answers, you can choose the second least frequent letter. Why? For example, if you notice there are only 4 Ds and you know the number of letters range from 5 to 16, there is a high chance that D will only appear 1 or 2 times.

code 0

	Sure	Rest	Expected	A	B	C	D
Number of certain answers	30.0	0.0	0.00	6.0	10.0	10.0	4.00
Number of the rest	0.0	20.0	5.00	8.0	4.0	6.0	1.00
Probability	0.0	0.0	0.25	0.4	0.2	0.3	0.05

Year 2018, code 0

PART D: Conclusions

- You should carefully re-check your answers if there are 4 or more of the same answer in a row.
- Across 50 answers, if there is a letter only appearing 4 times or there is a letter appearing 17 times, you're probably missing a letter or have 1 incorrect letter.
- In general, if you're able to do from 15 to 30 questions, choosing the least frequent letter is preferable (over sparsely choosing answers).
- If you're able to do more than 30 questions, choosing the **second** least frequent letter is preferable (over sparsely choosing answers).
- Using this strategy along with elimination method is a good idea. Simply by getting one over four eliminated, you've raised chance by 8.33% (from 25% to 33.33%)
- If you are totally clueless, you might want to consider using this above strategy because around 70% of the time, it will give you a better score. (The mean is 28%, the most optimistic number is 40%).
- **Again, this article focuses on a very specific case. In order to reach your goal, you will need to adjust the code.**
- **The code and the strategy won't work if you don't know the answer to at least 10 questions.**