When does Pooled Sampling save Money? - An Introduction to Modelling Pooled-Sample Testing

Aravinth Krishnan

March 2023

1 Abstract

The idea of Pool-sampling dates back to World War II. With logistical constraints on the availability of tests for particular diseases, a more resource efficient way to test soldiers was needed. So, the idea of pooling together a fixed number of samples and testing the pooled sample originated. As a more recent example, the World Health Organisation (WHO) declared COVID-19 a pandemic on March 2020. In certain regions of the world, there were shortages of testing supplies (the reagents and/ or plastic pipettes needed for the tests). Moreover, as students returned from remote learning to classrooms, many universities and public schools tested students weekly. The vast number of students being tested put a strain on budget as well as on testing supplies. In this case, pool sampling was helped ease the logistical and financial pressure. Other than testing for diseases, this technique is being used for drug testing programs in companies to help maintain a drug free work environment.

In deciding to pool-sample, key decision makers have to employ the most efficient approach to conduct it. In certain cases, pooled-sample testing might not even necessarily always be more efficient way to conduct tests compared to the traditional approach. When pool-sampling make sense, there might be different strategies to conduct pool sampling, such as considering the number of samples to pool together. Mathematical Modelling can be used to help people understand whether a particular strategy will save money in the long run.

This paper is made as reference for the Spring Mathematical Modelling Seminar in Kansas State University 2023 and mainly draws inspiration from COMAP's Mathematical Modeling Modules - Modelling Pooled-Sample Testing [Dav22]. Here, we will examine the strategy of pooling samples in order to save money on COVID-19 testing of students in a school or school system and readers will be taken through a series of open-ended questions and are encouraged to work through the problems before viewing the author's personal approach to the questions. Our goal is to find the strategy to save money in the long run using pooled-sample testing. This expose also serves as an introduction to the general theory/approach to mathematical modelling.

2 Introduction

Definition 2.1. *Sample pooling is the term used to describe the process of combining samples from two or more people. In pooling multiple samples:*

- *Mix a specificed number of individual samples together. If possible, save part of each sample in case more test must be done. If reserving a portion of individual samples cannot be done, further testing of the individuals may be required.*
- *Run one test on the pooled sample.*

Note that there are many ways to pool samples. Given n number of samples, we can pool them in pairs, or in 3s or in 10s. We will start with the simplest case, of mixing 2 samples. Starting with the simplest case allows us to generate a prototypical functioning model easily, and then we can use this as a base to increase the complexity of the case.

Next, we need to list all the assumptions we are going to make.

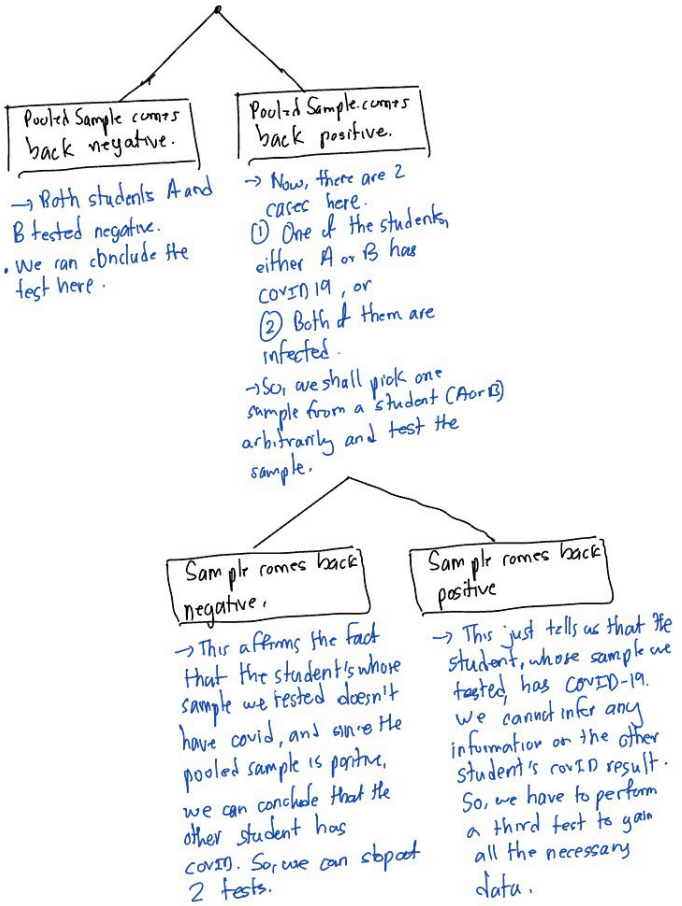
- Our first assumption is that the test are perfect. There are no errors in detecting the disease.
- We assume some of each sample can be reserved for future testing.

3 When can we save money in the long run by pooling 2 sample?

So, now we want to answer the question: When can we save money in the long run by pooling 2 sample?

3.1 1 pooled sample from 2 students

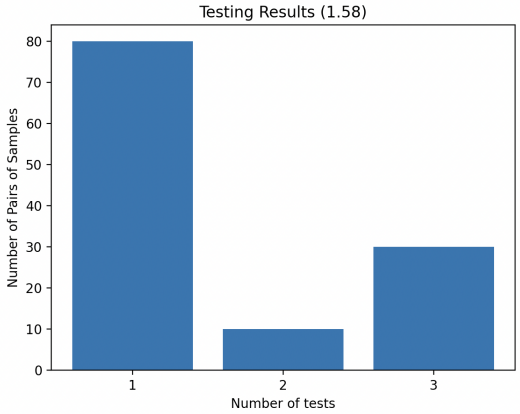
Firstly, let's consider 1 pooled sample from 2 students: What are the possible outcomes of the pooled sample? What do these outcomes imply on each individual student's infectedness? If the pooled sample tests positive, how do we determine which of the students would test positive for COVID? How many tests are needed in total for each different scenario?



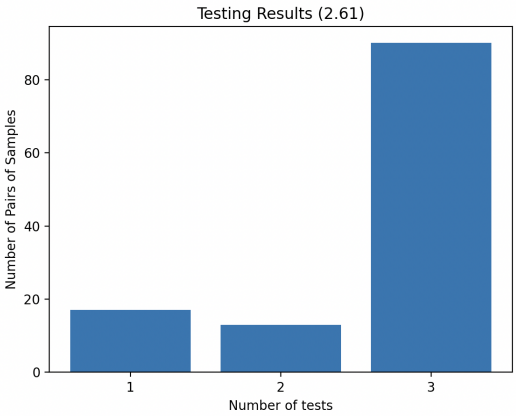
3.2 Pooled sample of pairs from n students

The questions above dealt with a single pooled sample of a pair of students. Now, given n number of students in the school, we want to keep a record of the number of test we did on each pair of the samples. Remember that we are trying to save money in the long run, so if each sample takes 3 tests, then we might as well traditionally, test each student once. So, what might be the best way to organise the data in a concise way?

We can compile the data in a bar chart. So, for example if the number of pooled samples that requires 1, 2 and 3 test are 80, 10 and 30, the bar chart will be as follows:



If the number of pooled samples that requires 1, 2 and 3 test are 17, 13 and 90, the bar chart will be as follows:



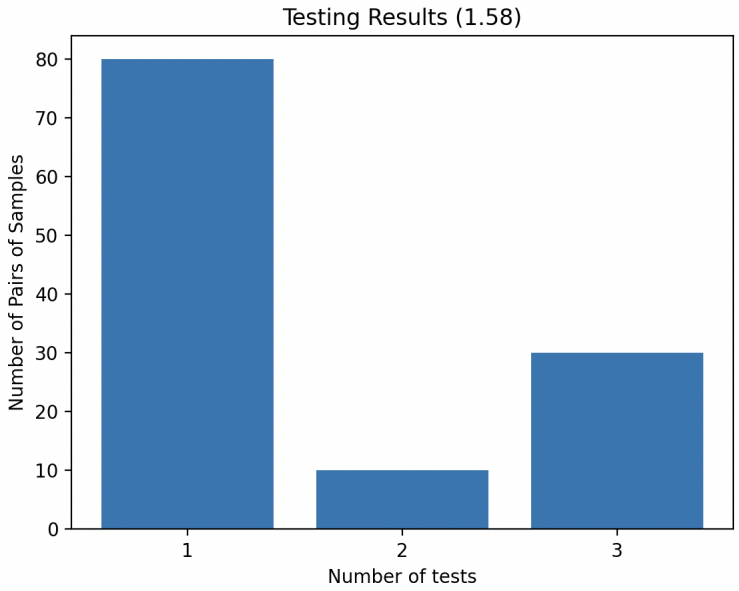
A Quick Brainteaser: If the number of sample pairs requiring one test is the same as the number of test requiring 3 test, does the sample pooling save money? Let's assume each test cost 30: Suppose that the total number of pairs needed testing is 120 and let the number of sample pairs requiring 1 and 3 tests be x each. Then, the total number of test required will be $x + 2(120 - 2x) + 3x = 240$, which is exactly the number of students there are. So, testing each student individually yields the same number of tests required. Thus, sample pooling does not save any money.

Definition 3.1 (Definition of Incidence)**.** *The rate of positive test in a population being tested is called incidence. Incidence can be expressed as a decimal or a percentage.*

Now, for each of the following situation, if you are using a strategy of pooling 2 samples, how many tests will you do on a pooled sample in each case. Will the school save money by pooling two samples?

- All the students have covid (but you do not know it). For this case, the incidence is 1.0 if expressed as a decimal.
- None of the students have COVID, and you do not know it. In other words, the incidence is 0.

Let's go back to one of the earlier scenarios.Suppose that the number of pooled samples that requires 1, 2 and 3 test are 80, 10 and 30, the bar chart will be as follows:



What is the average number of tests required per pool sample?

$$\frac{80 \text{ x } 1 + 10 \text{ x } 2 + 30 \text{ x } 3}{120} = 1.58$$

How can this value tell us whether the school has saved money? If the average number of test per sample pair is 1, this means that a total of 120 test were done. If the average number of test per sample pair is 2, this means that a total of 240 test were done. If the average number of test per sample pair is 3, this means that a total of 360 test were done. Note that if all the students were individually tested, a total of 240 test will be required. So, from this, the average of 2 sets the benchmark to decide if pooling-sampling the samples into pairs is a valid test for testing COVID-19. In this particular case, as the average $1.58 < 2$, pool-sampling will save money for the school. If the cost per test was 30, the school saves:

$$240 \text{ x } 30 - (1.58 \text{ x } 120) \text{ x } 30 = 1512. \tag{1}$$

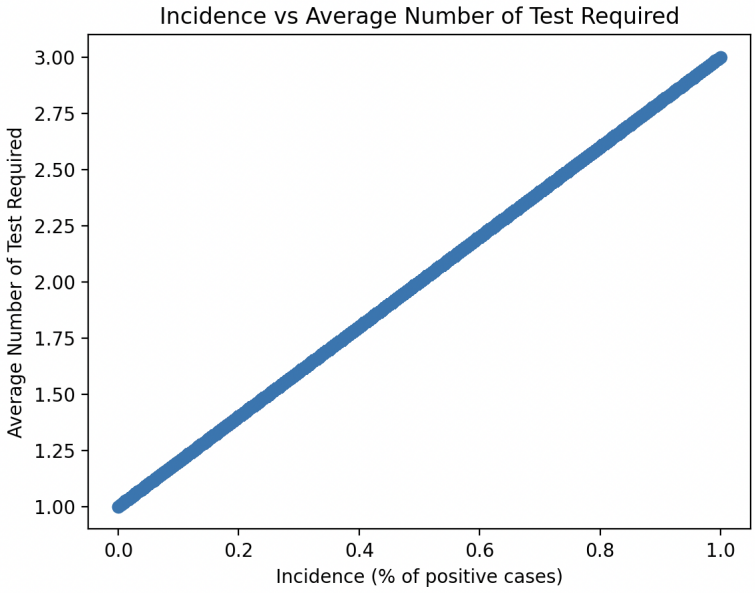
Now, make up our own values for the number of pooled samples requiring 1, 2 or 3 tests. The total number of pairs should be the 120. What are your values? Draw a bar graph for the values you chose. What was the average number of tests per pooled sample? From this data, did the school save money by pooling 2 samples? If each test cost 30, how much did the school save or lose by using the sample-pooling strategy compared to individual testing?

3.3 Relationship between Incidence and Average Number of Test Required

While working through these questions, you should have realised that incidence plays a major factor in determining the answer to this modeling question is the incidence of a COVID-19 in the school population. The next step is to use mathematics to describe a relationship between the incidence of COVID and the average number of test per pooled sample-in other words to create a mathematical model.

The author created a linear model using python to establish the relationship between incidence and the average number of test required for each incidence. To do so, the author generated all possible triples (x, y, z) which denote the the number of sample

pairs that requires 1, 2 and 3 testing respectively. As we deduced earlier, each of the cases above correspond to $0, 1$ and 2 students having COVID-19. From this, since we know the total number of students, we can calculate the both the incidence and average number of test required for that particular case of (x, y, z). These tuple of data was then stored in an excel sheet (available here: add in link here). From this set of data, we can generate the linear model (The code to access the calculations can be accessed here:) :



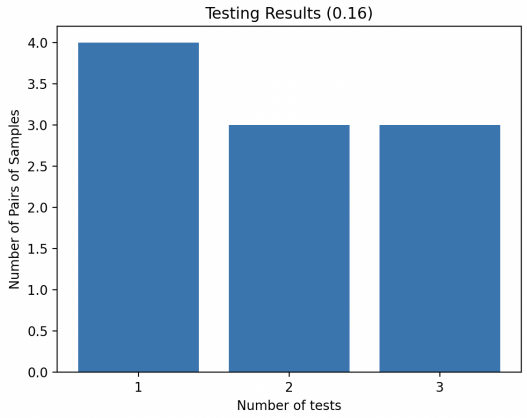
As we can see, there is a direct correlation between incidence and the average number of test required. The correlation is almost 1. Earlier, we deduced that the average of 2 sets the benchmark to decide if pooling-sampling the samples into pairs is a valid test for testing COVID-19. So, as long as the incidence is known and is lesser than or equal to 0.5, pooling the sample into pairs saves resources, both logistically and financially.

So, we have created a linear model for predicting the average number of tests per sample in pooled, two-sample testing given the incidence of COVID in a population. We won't be able to test our model directly with real-world data. Instead, we will be simulating the testing situation by using programming.

Suppose we are simulating a situation in which 50% of the school population would test positive for COVID. We will use the random number generator in python, from the random module, to randomly generate a number between 0 and 1, where 0 indicates that a student is COVID-negative and 1 denotes COVID-positive. The python program simulates the following process:

- Each student in the pair-testing group will be randomly given an integer 0 or 1.
- The two tested students from each group should tell the tester whether their pooled sample is positive or negative. (Recall that a pooled sample is positive if at least one student in the tested pair tests positive.)
- If the pool sample is positive, the tester should use a random device to pick one of the students from the pair to test first. For the test, that student should report their test result to the tester. If the selected student's test is negative, then no additional tests are required. If the selected student's test is positive, then the second student in the pair must be tested.

For a first try, we suppose that there are 10 pairs of students. We record a summary of times one, two and three tests are required from the 10 pooled samples and record the results in a barchart. One example includes:



The python console output looks as follows:

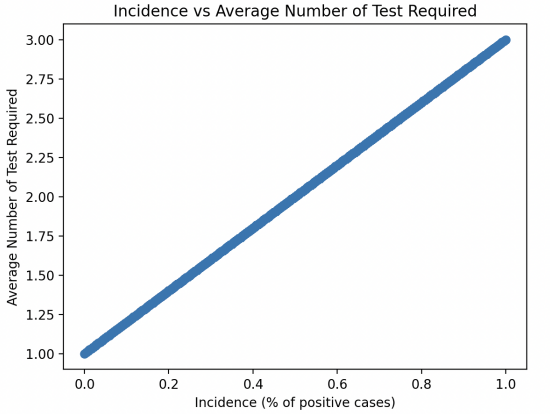
```
[[0, 1], [0, 0], [0, 0], [1, 0], [0, 0], [1, 1], [0, 0], [1, 0], [0, 1], [1, 0]]
['Positive', 'Negative', 'Negative', 'Positive', 'Negative', 'Positive', 'Negative', 'Positive', 'Positive', 'Positive']
[2, 1, 1, 2, 1, 3, 1, 2, 3, 3]
```

We can see that the average number of tests for this round of simulation is 0.16. Comparing the results from the linear model we constructed, we see that the simulation cast doubts on the prediction from our previous model as the value of the simulation is lower than that of the linear model.

We shall perform another simulation in which the incidence of COVID in the school's population is 0.5. This time we are going to use a much larger number of testing-pairs, $10,000$. 2 of the outcomes of the simulation are as follows:

The simulation done on 10,000 sample pairs is more reliable that the result obtained from 10 pairs. Both results cast doubts on the linear model, which predicted the average number of tests to be 2 when the incidence is 0.5. We created the linear model under the assumption that the relationship between incidence and the average number of tests required is linear. The simulation shows that this is not necessarily the case.

In all of the cases above, we assumed that a positive test result from the pooled sample was followed by testing one of the students in the pool for COVID, and if neededm testing the second student. In Massachusetts, the pooled-sample testing protocol for schools used a different strategy. If the pooled sample tested positive, then all the students in the pooled sample were tested individually using a Rapid COVID test. With this change, our linear model becomes:

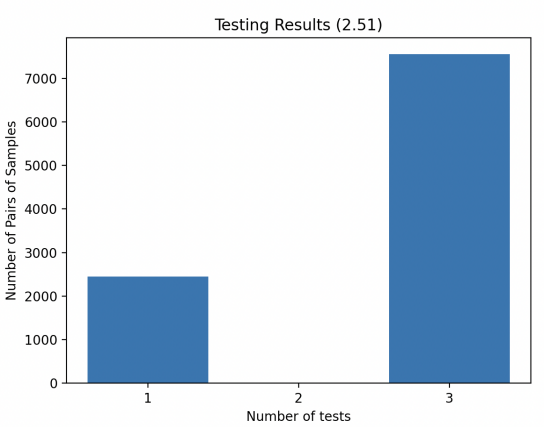


So, we see that the linear model doesn't change.

However, in the simulation for 10,000 pairs, the following change occurs:

- If the pooled sample is negative, only one test is needed.
- If the pooled sample is positive, then both students are tested. Three tests are needed.

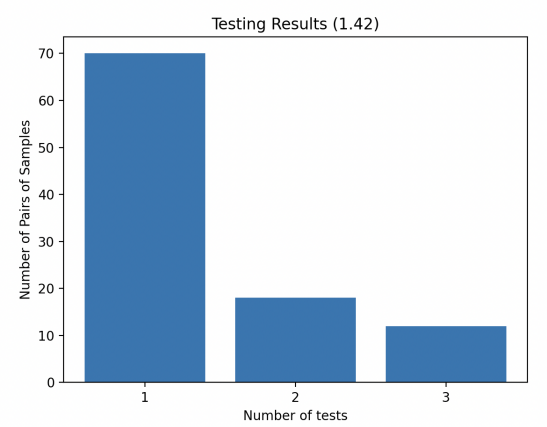
In this case, either one test is performed or 3 tests are performed. So, an example of a simulation output is as follows:



In the previous activities, we used simulated data to test the extremes model created. The results from the simulations cast doubt on this linear model, which was based on the incidence extremes of 0% and 100%. Now, we will collect simulated data and then use that data to build a better model. Remember that the original goal is to use the model to answer the question: Over the long run, when can you expect to save money by pooling two samples?

To simulate a situation in which 50% of the students have COVID, we can flip a coin. To simulate a situation in whuch the percentage is not 50, coin flipping doesn't work. Drawing colored chips out of a container is a btter option. For example, to simulate a situation in which 30% of the students have COVID, we can use 10 chips, some of which are red, and some of which are blue. Let red denote a positive COVID test. So, there should be 3 red and 7 blue chips in the container. If the pool sample comes back positive, the tester randomly selects one of the students from the pool. The selected student should draw a chip from the container to determine their COVID status. If this student tests positive (the chip drawn is red), the second student needs to be tested.

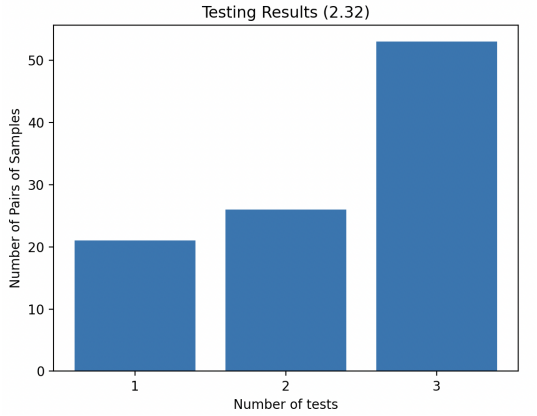
To simulate COVID testing using calculator and computer programs, a decimal between 0 and 1 can be randomly generated. Suppose that the 30% of the population has COVID. Then, if the random decimal is below 30%, then the program counts the person as testing positive. If the random number is above 30%, the program counts the person as testing negative. Now, let's use the simulation to determine the average number of tests if the incidence of COVID is: $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. For each incidence, we let the number of trials be 100:

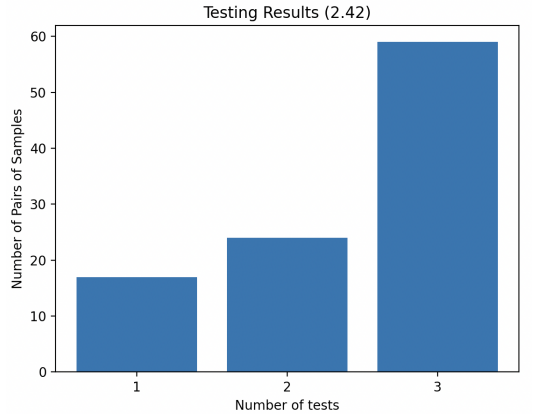

Testing Results (1.42)
Number of Pairs of Samples
0
10
20
30
40
50
60
70
1
2
3
Number of tests

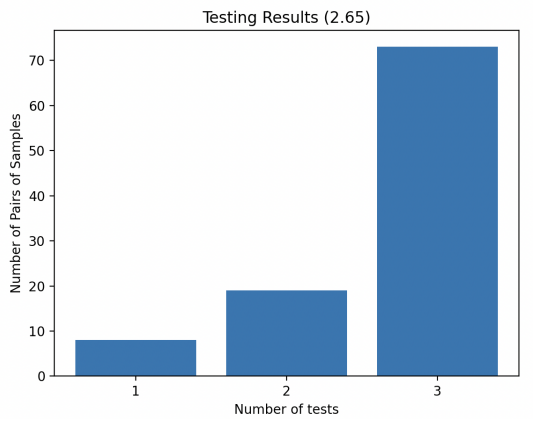

Testing Results (1.48)
Number of Pairs of Samples
0
10
20
30
40
50
60
70
1
2
3
Number of tests

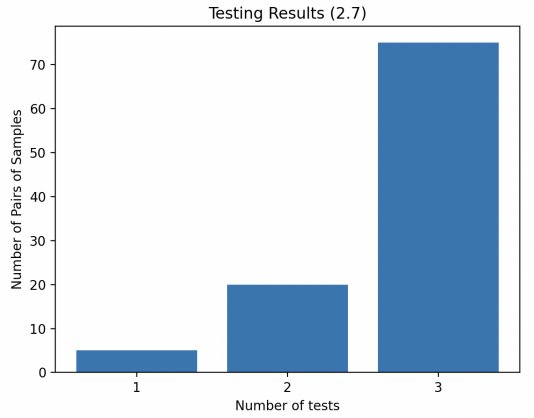

Testing Results (1.72)
Number of Pairs of Samples
0
10
20
30
40
50
1
2
3
Number of tests

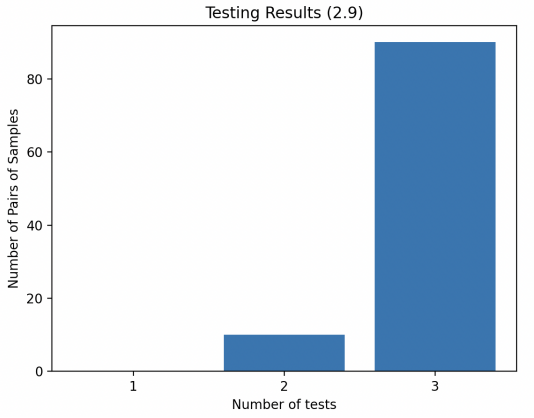

Testing Results (1.98)
Number of Pairs of Samples
0
5
10
15
20
25
30
35
1
2
3
Number of tests

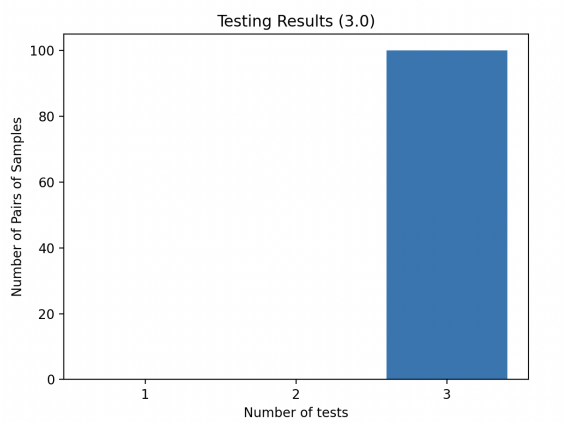

Testing Results (2.32)
Number of Pairs of Samples
0
10
20
30
40
50
1
2
3
Number of tests


Testing Results (2.42)
Number of Pairs of Samples
0
10
20
30
40
50
60
1
2
3
Number of tests


Testing Results (2.65)
Number of Pairs of Samples
0
10
20
30
40
50
60
70
1
2
3
Number of tests


Testing Results (2.7)
Number of Pairs of Samples
0
10
20
30
40
50
60
70
1
2
3
Number of tests

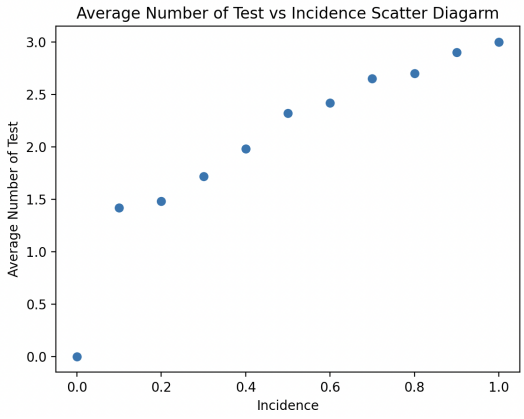

Testing Results (2.9)
Number of Pairs of Samples
0
20
40
60
80
1
2
3
Number of tests



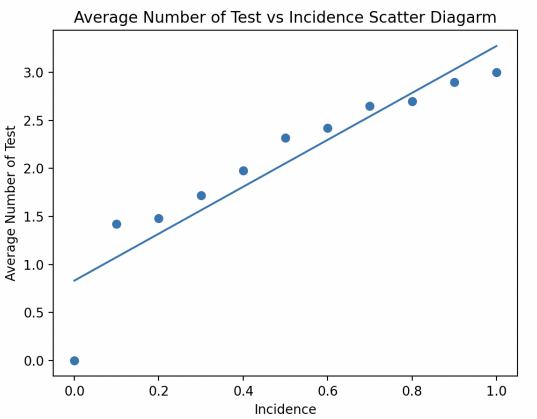
We can consolidate this data on a table as follows:

	Number of tests			
Incidence	1	2	3	Average Number of Tests
0	100	0	0	1
0.1	90	10	0	1.42
0.2	75	20	5	1.48
0.3	73	19	8	1.72
0.4	59	23	18	1.98
0.5	52	26	22	2.32
0.6	37	25	38	2.42
0.7	23	23	54	2.65
0.8	16	16	68	2.7
0.9	13	17	70	2.9
1	0	0	100	3

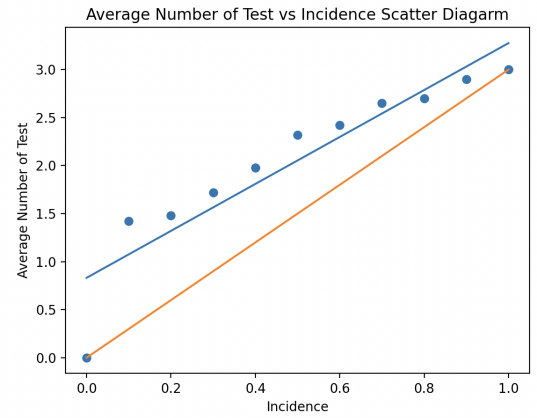
and a scatterplot of the data:



Now, let's fit a regression line into the scatter diagram to see the relationship between the incidence of COVID and the average number of test required:



To compare the regression line to our linear model:

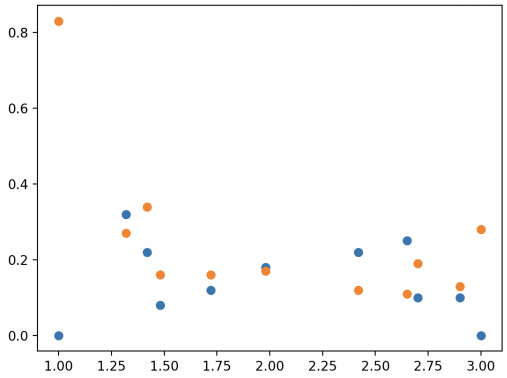


Notice that the linear regression line appears to summarize the pattern in the data than the line that connects the extreme points. For the linear model connecting the extreme points, all the other points lie above the line. Most of the points in the scatterplot lie closer to the regression line than to the line connecting the extreme points. However, even though the scatter plot is a nicer approximation, both are inadequate to describe the pattern of the dots in the scatter diagram.

We will now introduce an important tool to evalaute a model - a table or a graph of residuals. The residuals are the difference between the average number of tests (calculated from the simulated data) and the average number of test predicted from the model. Calculate the residuals for each of the linear models graphed in the scatter plot:

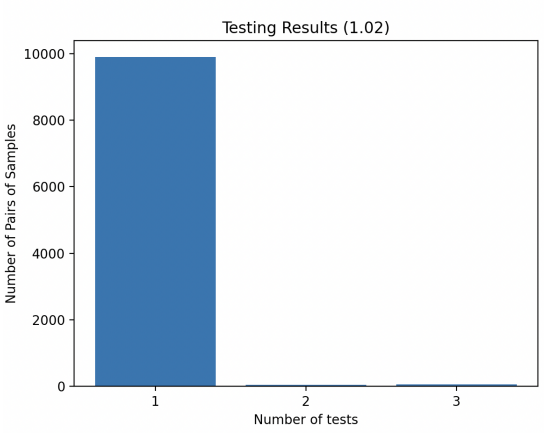
Incidence	Average Number of Tests from Simulation	Linear Model Predicted Value	Linear Model Residual	Regression Model Predicted Value	Regression Model's Residual
0	1	1	0	0.83	0.83
0.1	1.42	1.2	0.22	1.08	0.34
0.2	1.48	1.4	0.08	1.32	0.16
0.3	1.72	1.6	0.12	1.56	0.16
0.4	1.98	1.8	0.18	1.81	0.17
0.5	2.32	2	0.32	2.05	0.27
0.6	2.42	2.2	0.22	2.3	0.12
0.7	2.65	2.4	0.25	2.54	0.11
0.8	2.7	2.6	0.1	2.79	0.19
0.9	2.9	2.8	0.1	3.03	0.13
1	3	3	0	3.28	0.28

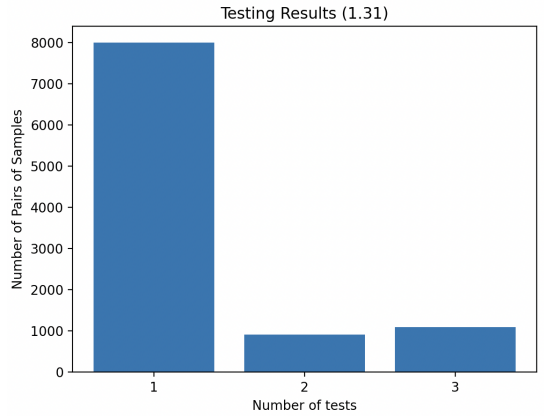
A residual plot, a plot of the residuals versus the independent variable (in this case, incidence), is a good diagnostic tool for deciding whether a model is adquate to describe the pattern in a scatterplot. A residual plot in which the dots appear randomly scattered with no strong patterns and (though less important) a fairly even mix of dots above and below the horizontal axis indicates that the model is adequate to describe the pattern in the data:

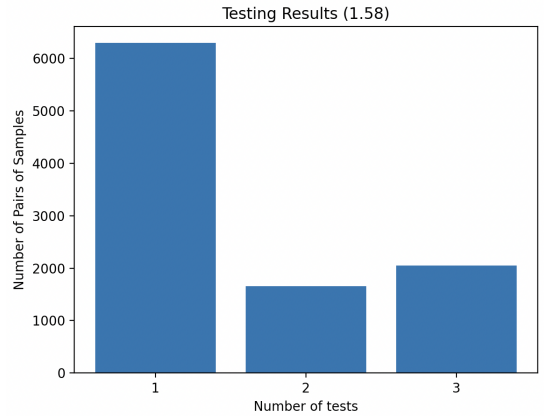


The residuals from the regression model have a better balance between positive and negative values from the extreme model. However, note that our residual plot shows a strong curved pattern. The linear model was not able to capture the curved pattern in the data. So, we need a different model - an equation that captures the curve in the shape of the data.

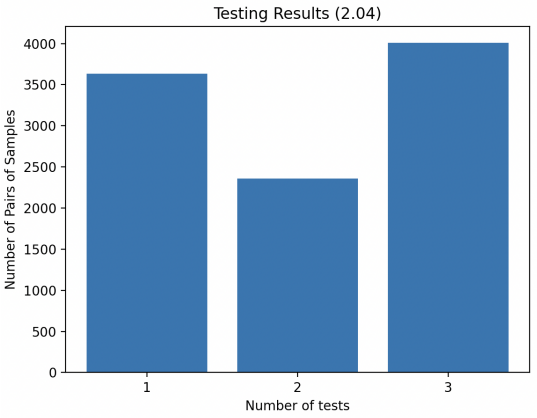
Our simulation previously involved only 100 trials. Using simulated data based on more trials will give better results. So, let's simulate data based on 10000 trials for the same indicience level as done for 100 trials:

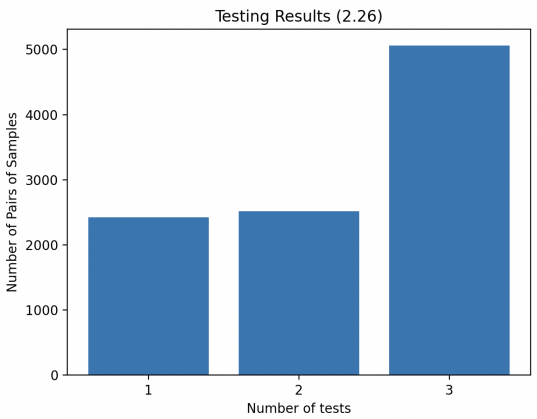

Testing Results (1.02)
Number of Pairs of Samples
Number of tests
0
2000
4000
6000
8000
10000
1
2
3

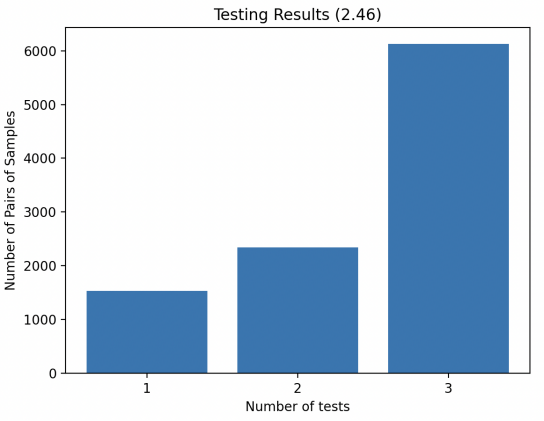

Testing Results (1.31)
Number of Pairs of Samples
Number of tests
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3

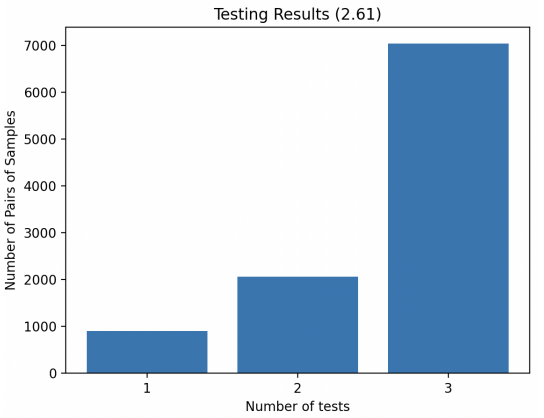

Testing Results (1.58)
Number of Pairs of Samples
Number of tests
0
1000
2000
3000
4000
5000
6000
1
2
3

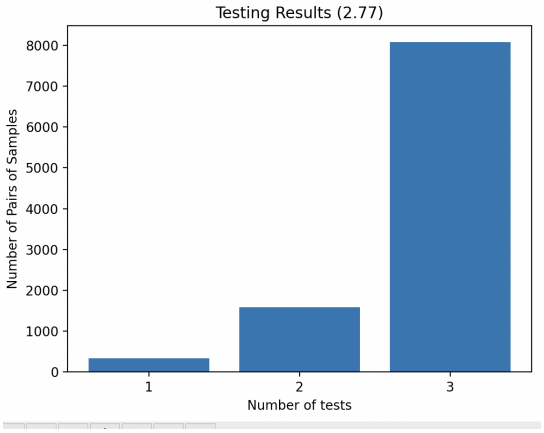

Testing Results (1.82)
Number of Pairs of Samples
0
1000
2000
3000
4000
5000
1
2
3
Number of tests

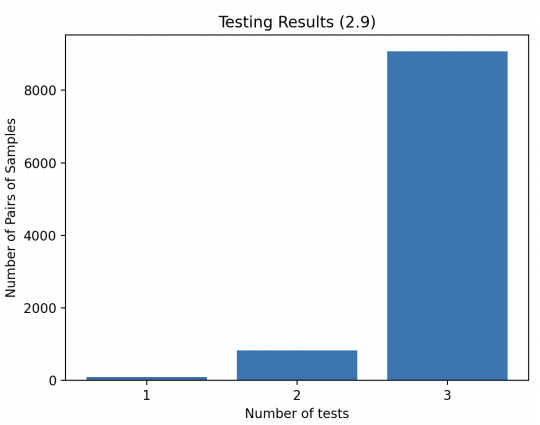

Testing Results (2.04)
Number of Pairs of Samples
0
500
1000
1500
2000
2500
3000
3500
4000
1
2
3
Number of tests

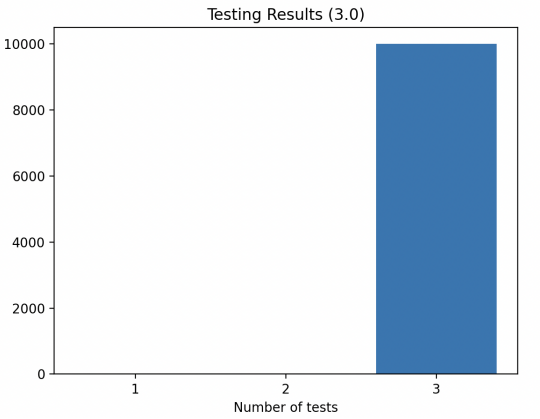

Testing Results (2.26)
Number of Pairs of Samples
0
1000
2000
3000
4000
5000
1
2
3
Number of tests


Testing Results (2.46)
Number of Pairs of Samples
0
1000
2000
3000
4000
5000
6000
1
2
3
Number of tests


Testing Results (2.61)
Number of Pairs of Samples
0
1000
2000
3000
4000
5000
6000
7000
1
2
3
Number of tests


Testing Results (2.77)
Number of Pairs of Samples
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
Number of tests

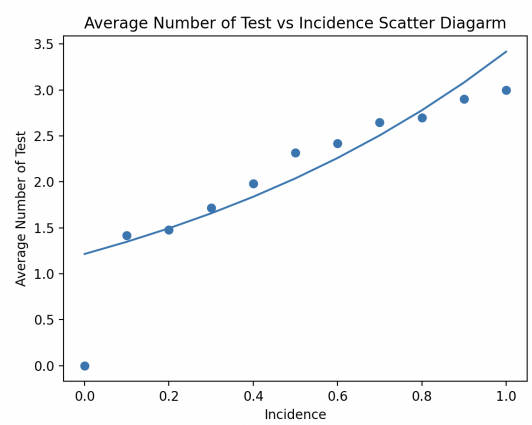


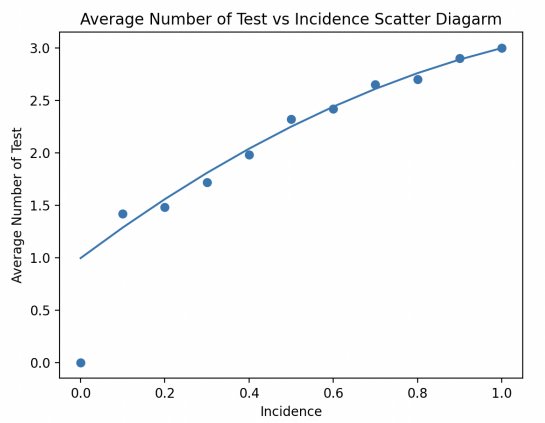


Once again, we can consolidate this data on a table as follows:

	Number of tests			
Incidence	1	2	3	Average Number of Tests
0	9998	1	1	1.02
0.1	8000	1000	1000	1.31
0.2	6500	1500	2000	1.58
0.3	4900	2000	3100	1.82
0.4	3600	2400	4000	2.04
0.5	2400	2600	5000	2.26
0.6	1500	2700	6300	2.46
0.7	900	2100	7000	2.61
0.8	200	1800	8000	2.77
0.9	200	500	9300	2.9
1	0	0	10000	3

The curvature we describe earlier motivates us to explore using the exponential and quadratic regression model. The models we obtain are as follows:





From the scatter diagrams, we see that the quadratic model appears to hug the points in the scatterplot. The exponential model is concave up while the pattern of the dots in the scatterplot forms a concave down pattern. So, based on the quadratic model and the fact that money is saved as long as the average number of test required is less than 2, as long as the incidence is less than 40%.

Bibliography

[Dav22] Marsha Davis. *Modeling Pooled-Sample Testing, Part I.* Vol. I. A COMAP Modeling Module. Consortium for Mathematics and its Applications, 2022.