Optimized Pool Testing for COVID-19 with focus on the Indian Demographic

Prajval Koul*

Abstract:

The COVID-19 pandemic has gripped the entire globe in a span of less than half a year claiming a recorded 200,000 lives [1] till date. All the countries are going for rapid testing aiming to hopefully contain all the carriers before it takes the form of an epidemic, or what the populous calls Stage-4. The current method of testing used by the governments has been compared to the method proposed. Towards the end, some figures have been calculated to show the actual capital saved by using the proposed method.

Motivation:

The proposed solution was developed within the four walls of the writer's homes during the lockdown in India. The current COVID-19 contamination has been described by an analogous problem for simplicity of explanation. Given a set of identical looking beakers, each containing either acid or water, and some empty ones as well, what is the minimum number of litmus strips (non reusable) required to tell apart the nature of the contents of each beaker, given that a certain percentage of beakers contain acid? This problem has been described in comprehensive details in the following sections. It should be noted that this problem is just for demonstration purposes. The following sections discuss the actual COVID-19 pandemic and how this mock problem can be used to understand the actual testing scenario.

Description of the problem:

As stated before, consider the following problem. There are B beakers containing either acid or water. All the beakers are indistinguishable when viewed from the outside. It is given that P percent of the beakers contain acid. We have to calculate the minimum number of litmus strips required to correctly label all the beakers. It would be customary to report the corresponding algorithm used to do the same.

This can directly be related to the current scenario. The beakers can be thought of as the subjects, the acid can be thought of as the virus and the litmus strips can be thought of as the testing kits, which are also non reusable.

One simple answer would be to use B number of litmus strips, one for each beaker. That can be taken to be the worst case count i.e. the number of strips should never cross B.

* Department of Computer Science and Automation, Indian Institute of Science, Bangalore, Karnataka, India (560012).

Solution and Related Discussion:

Consider a simple variant of the problem. Given that we have only 1 beaker with acid in the set, how many litmus strips will be required to locate it. This can be found out using divide and conquer approach. We divide the beakers into 2 equal sized subgroups, pool the samples of all the beakers in each group, and test them using a pair of litmus strips. Since there is only 1 acid containing beaker, only one of the 2 subgroups will test positive. Now the same procedure is applied recursively on the subgroup which tested positive in the previous step. This procedure is repeated until we find the beaker containing acid.

Now, given *B* beakers, the time complexity of the above problem is given by the following relation:

$$T(n) = T(\frac{n}{2}) + 2$$

The solution of this relation is $T(n) = 2\log_2(n)$ ($O(\log(n))$).

This is a very promising landmark to move forward.

Now, if we are able to divide the initial set of beakers into subgroups such that each subgroup contains an expected quantity of 1 acidic beaker with respect to the overall set, we can use the above algorithm on each subgroup to obtain an optimized count.

Let's consider for a moment that the above division has been made somehow. Now, if we pool the contents of a subgroup and it tests negative, we don't have to proceed further with the initial algorithm on that subgroup. Hence, we can think of the worst case to be having exactly 1 acid containing beaker in a subgroup. Let's go with this case only, i.e. each subgroup divided has exactly 1 acid containing beaker.

Now, we are given that P percent beakers contain acid. This means,

Number of acid containing beakers = $\frac{B*P}{100}$

i.e. if we make subgroups with size being the ratio of total number of beakers to the number of acid containing beakers, we get,

$$\text{Group size} = B \Big/ \frac{B*P}{100} = \frac{100}{P}$$

If we divide the set of beakers uniformly according to the above size whilst considering the worst case scenario, we will arrive at a scene where in each subgroup, we will have exactly 1 acid containing beaker. After that, we can carry out the final solution.

Minimum number of litmus strips required in the worst case is given by the expression,

$$2\,\frac{B*P}{100}\log_2(\frac{100}{P}) \qquad \ldots\ldots\ldots\ldots(1)$$

It should be noted here that we didn't add another $\frac{B*P}{100}$ to this expression, which corresponds to the number of strips required to test the initial pools for positivity. This is actually a redundant activity because even

if we don't test these pools, eventual testing from the groups using the divide and conquer approach will yield correct results.

Consider, for instance, a simple case to support this claim. There are 8 beakers and 25% of them are acidic. We divide the beakers into $\frac{B*P}{100}$ groups, which comes out to be 2, in this case (for worst case). Now, we don't test these pools. We simply proceed further. Each subgroup of 4 beakers is broken down into sets of 2, pooled and tested. The worst case calls for a positive and a negative result. What if the entire group was negative? One would argue that we could've told this by using a single strip on the entire pool rather than 2 strips in the subgroups. This claim is true to some extent only. A counter argument would be something like; if a group, or a pool in the first stage, has a dense quantity of acidic beakers, i.e. it has more than 1 acidic beakers, the initial test on the first level proves to be redundant because we will have to use another strip to cater for the extra beaker that was not taken into consideration. This is a case little divergent from the worst case. Hence, the extra strip which was thought to be 'wasted' was actually compensated for in the other pool. Hence, the testing at the initial stage is redundant, i.e. the extra $\frac{B*P}{100}$ is not to be added in expression (1).

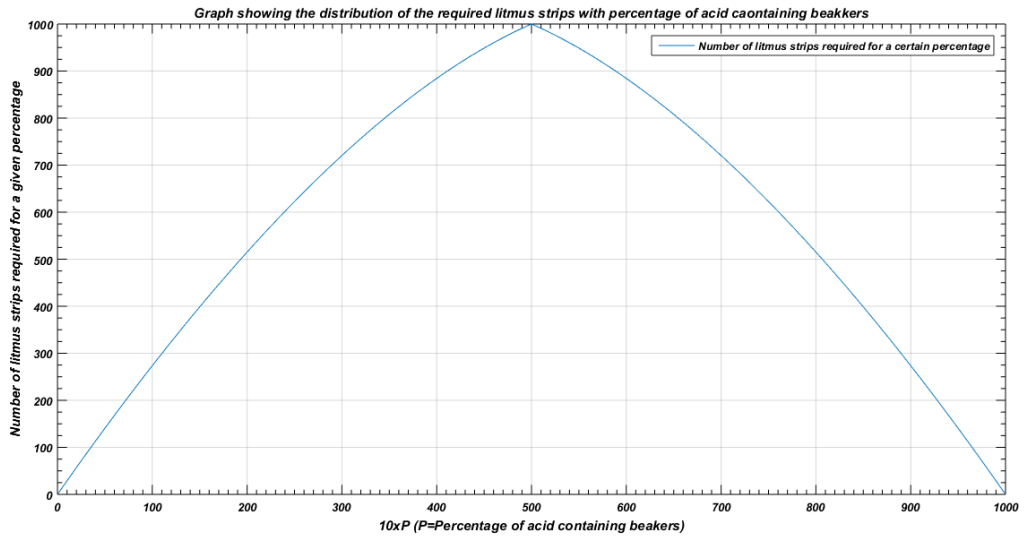

Figure 1: Graph showing the number of litmus strips required for 1000 beakers with respect to the percentage of acidic beakers (considering the mirror image concept).

Now, although the expression (1) obtained above is not symmetric about any axis (in general), a simple trick has been employed. The first half of the curve is replaced by a mirror image of the second half about the line $x = \frac{1}{2}$ (assuming percentage goes to 1, for simplicity). This is because searching for *k*% positive cases is same as searching for (100-*k*)% negative cases and vice versa. Figure 1 shows this modified plot for the previous expression (1) for $B = 1000$. Replacing the second half with the reflection of the first half is not economical because the rate of growth of the first half of the expression is greater than rate of decay of the second half, and hence doing the first alternative is better. Notice that the expression evaluates to B when $P = 50\%$.

Now, the current testing scenario (followed by governments of many countries) has been a little different than the one described here. Let's stick to the analogy used. Considering *P* percent beakers to contain acid, the beakers are divided into *G* groups. The samples from the members of each group are pooled and tested for acid. If a pool tests positive, then all the beakers in that group are tested individually. What should be the optimized value of *G* so as to minimize the number of litmus strips used?

The normal solution adopted is to fix a certain value of *G* at random, say 5, and go with the procedure. This way is highly inefficient. To check the viability of this method, let's try to optimize *G*.

Again we assume worst case scenario, i.e. each group has one acidic beaker (for understanding purposes). If $\frac{100}{P}$ (Group size from proposed method) is less than $\frac{B}{G}$ (Group size if we have *G* groups), then we have to test all the beakers because, by pigeon hole principle, each group will have at least one acid containing beaker per group. In this case, we straightaway take *B* strips and start testing, because going by this way would take *B* strips and *G* more strips, which is undesirable, since the number of strips should never cross *B*. The other case is when $\frac{100}{P}$ is greater than $\frac{B}{G}$. In this case, we need *G* strips for *G* number of groups and $\frac{B}{G}$ strips for each of the $\frac{B*P}{100}$ groups that test positive, in the worst case. Hence, in the second case the number of litmus strips is given by the expression

$$G + \frac{B*P}{100}\ \frac{B}{G} \quad \text{.....(2)}$$

Using simple calculus, we optimize the expression with respect to *G* and obtain,

$$G = \frac{B\sqrt{P}}{10}$$

Putting this value in expression (2) we get,

$$L = \frac{B\sqrt{P}}{5}$$

This quantity is more that the quantity returned by expression (1) for all values of percentage less than ~93% (or greater than ~7%, by the mirror image concept, which means that the quantity returned by expression (1) is optimal for $P \in [7,93]$). Now, one would argue that we can break the function obtained by the proposed approach. But we don't need to do that because the percentage of cases has crossed the ~7% mark long time back.

This section hence buttressed the efficiency and optimality of the proposed approach when compared to the approach currently followed.

Logistics for transplanting the solution to the domain of COVID-19:

As discussed in the previous section, the proposed solution can directly be related to the current scenario. The beakers can be thought of as the subjects, the acid can be thought of as the virus and the litmus strips can be thought of as the testing kits, which are also non reusable. Some additions have to be made in order to render the solution practical and implementable.

An initial estimate of *P* can be made out by testing a statistically random sample. This will provide an approximately correct estimate for the percentage of infected population, given instant sample collection. Since that is not feasible, there are 2 remedial ways to compensate for that. The first one includes collecting and testing the random sample set (which is very small) after all the samples have been collected. This way, one will get the current estimate of *P* correctly. The second one employs regression to predict the value of *P* for the moment testing starts. This alternative, although numerical, will yield results for exact moments, which will help enhance the overall accuracy.

Let us consider the regression approach for a moment because that can use some numerical analysis. The first method is a simple mechanical modification. To verify the regression approach, current data has been manually acquired by daily analysis of case count in India.

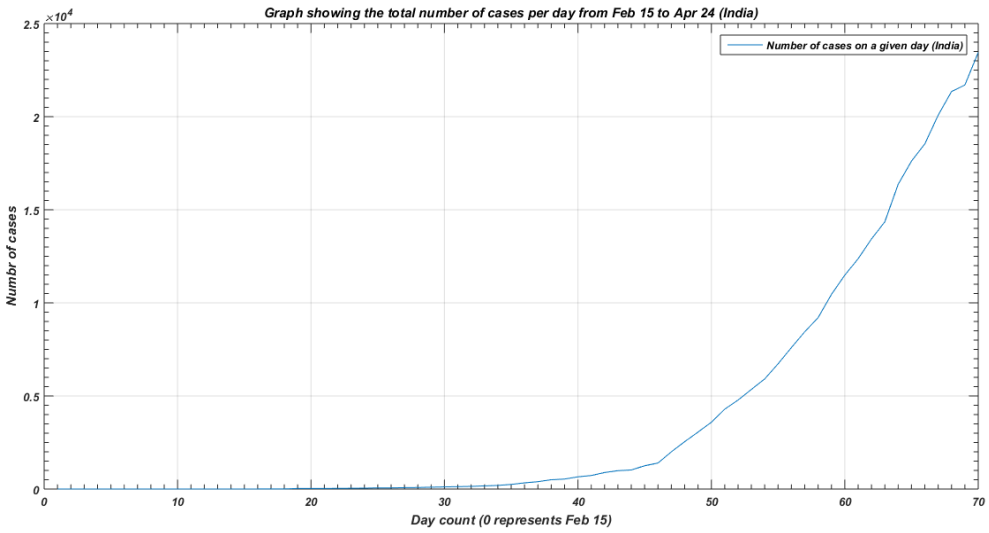

Figure 2: Curve showing total number of cases on each day from February 15 to April 24 in India.

It is broadly known that the positive cases grow exponentially in number. Figure 2 shows the distribution of case count in India for the period February 15 to April 24. Hence, an exponential curve is fit upon the data distribution to predict the future rates of contamination. This analysis is not meant for extensive prediction and advocating the accuracy of any model, rather it is aimed at addressing the current situation of emergency. Using the non linear least square regression, root mean square error (RMSE) is used to calculate the coefficients for the proposed exponential curve (Refer to appendix for more details). For demonstration purposes, data of previous *N*-1 days was used to predict the number of cases on the N^{th} day. Figure 3 shows the error values for each day using data from previous days and figure 4 is a comparison of the actual case count to the fit curve.

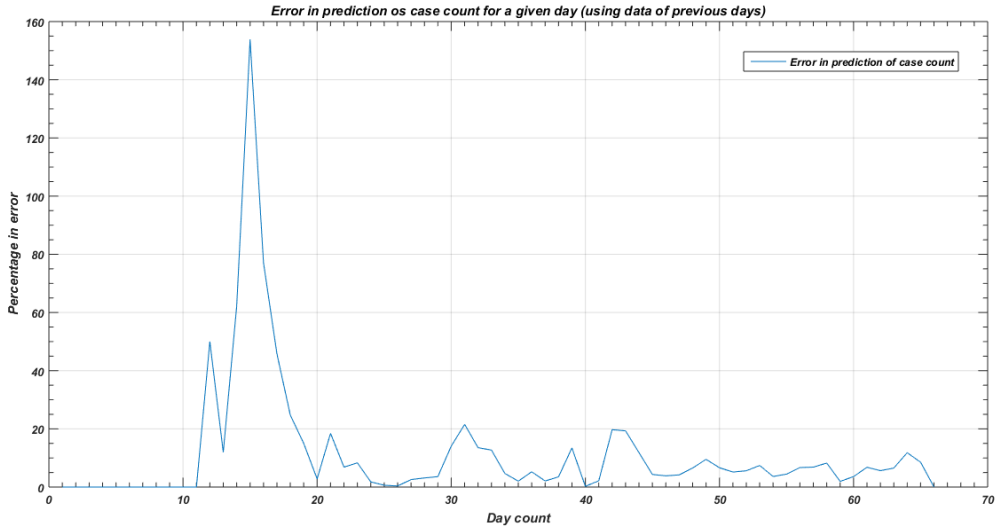

Figure 3: Percentage error in prediction of cases for each day using least square regression.

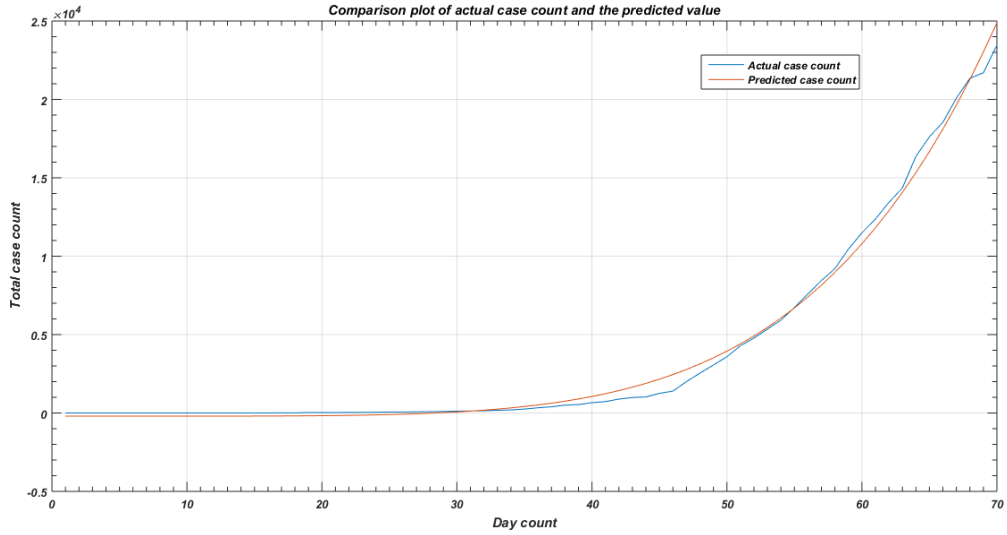

Figure 4: Exponential curve fit over the case distribution for the timeline mentioned in the text.

Another thing to be kept in mind is the success rate of the kits. Let the success rate be *S*. This means, if we use *B* kits, *S***B* number of kits will return correct results, be it positive or negative. This helps specify an error bound on the estimated number of cases.

If there are *K* actual cases overall, then the tolerance limits for the reported cases would be $K \pm S\%$. To over-approximate the result, we take the upper bound using the tolerance limits. This means, we take the reported cases as the lower bound value (*K* - *S*%) and then calculate the upper bound value (*K* + *S*%) to estimate the total cases. This might slightly increase the estimated case count, but this count, being an over-approximation, is safe when the subject of containing the carriers is concerned.

As pointed out earlier, this study has been done with a special focus on the Indian demographic. The estimated population of India is a massive 1,350,000,000. The least square regression returns $P = 12.4$ (Refer to the appendix section for details). Thus, we get an estimated 766,570,000 kits required. Assuming a 29% error rate per kit [2], 1,211,180,600 kits will be required. This quantity is, as expected, less than the overall population value.

Another thing to be considered, from a pathological point of view, is, whether pooling a large number of samples causes excess dilution (or some other pathological constraint) [3].

Impact on capital:

As seen in the previous sections, the expected number of kits saved is 138,819,400 - 583,430,000. This count is significant when the overall population is taken into consideration. We can find the expected savings by using a cost estimate value per kit. The price distribution is shown in figure 5. Note that the distribution remains similar to that shown in figure 1.

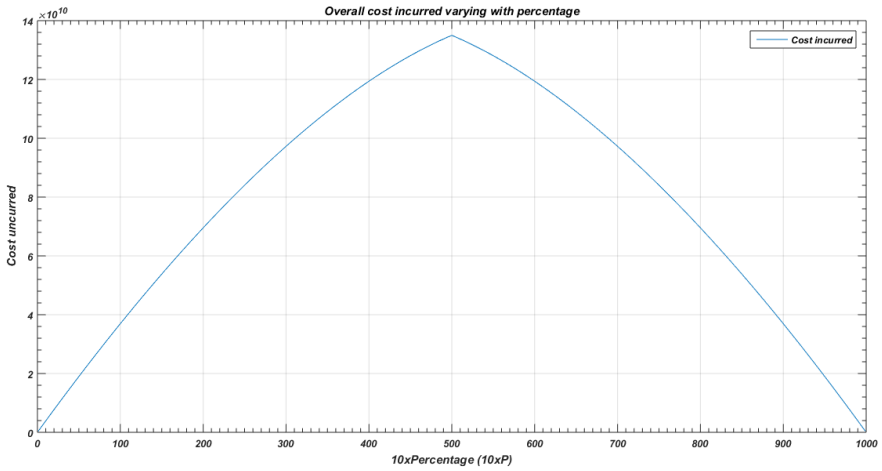

Figure 5: Distribution of total price varying with percentage, for a population of 1,350,000,000.

Consider the scenario in India (the primary focus of this study). A rapid testing kit costs INR 1000 (or $35 - $50 in the US, varying across states). So an estimated INR 138,819,400,000 - 583,430,000,000 can be saved by implementing this strategy, which is a huge amount, given such emergency like situations.

Conclusion:

The above numerical analysis paves a way for a successfully implementable rapid pool testing technique which will save a lot of capital investment whilst helping contain the carriers and hence the pandemic.

References:

[1]https://www.worldometers.info/coronavirus/

[2]https://www.hindustantimes.com/india-news/coronavirus-update-covid-19-rapid-testing-kits-use-suspended-for-two-days-due-to-errors/story-iUw9gw1CL2KL5BwmApQ5sN.html

[3]https://icmr.nic.in/content/covid-19

[4]https://www.statista.com/statistics/1107186/india-coronavirus-covid-19-testing-numbers-by-state/

[5]https://www.business-standard.com/article/health/covid-19-are-we-testing-enough-here-s-how-india-fares-versus-others-120040901595_1.html

Appendices:

Actual Data from India during the period February 15 to April 24:

Following is the daily count of the total number of cases in India from March 21 to April 24. Sources have been cited in the previous sections. The values have been arranged sequentially according to the calendar.

February 15 – February 29:

3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3.

March 1 – March 31:

3, 6, 7, 29, 30, 31, 34, 40, 47, 62, 62, 74, 82, 100, 114, 129, 143, 169, 194, 249, 332, 396, 499, 536, 657, 727, 887, 987, 1024, 1251, 1397.

April 1 – April 24:

1998, 2543, 3059, 3588, 4289, 4778, 5351, 5916, 6725, 7600, 8446, 9205, 10453, 11487, 12370, 13430, 14352, 16365, 17615, 18539, 20080, 21352, 21700, 23452.

The corresponding curve is shown in figure 2.

Regression Analysis:

Non linear least square regression, as mentioned earlier, was employed to fit an exponential curve on the above distribution.

The curve $a.x^b+c$ was fit over the distribution and the following parameters were obtained.

$a = 3.36x10^{-6}$

$b = 5.35$

$c = -202.5$

$SSE = 1.438x10^{7}$

$RMSE = 463.2$

Now, the estimated case count for April 25 according to the above analysis comes out to be 24883 to be discrete. Hence, out of a tested quantity of ~200,000 [4-5], the contamination percentage comes out to be 12.44%.