Some Remarks on Group Testing for COVID-19.

Michael R. Pilla

June 30, 2020

Abstract

The most effective way of controlling COVID-19 outbreaks is to test everyone on a regular basis. The problem, of course, is that this immensely costly and inefficient. In this paper, we discuss utilization of the method of group testing instead of individual testing, look at some results related to COVID-19, and discuss the problem of false negatives.

1 Introduction

Whether it is a college campus or a local town, the best way to control COVID-19 and keep the curve flattened, is to detect outbreaks as early as possible. This requires a significant amount of testing. The more testing done, the better chance we have of catching an outbreak before it spreads. The ideal situation would be to have the capacity to test everyone on a regular basis, a request too steep to be reasonably met given individual costs and test availability. But this is not the only option we have. Another strategy is to group individuals together, combine their samples into one and test this combined sample. This is known as group testing. The idea of group testing can be traced back to a specific paper written by Robert Dorfman in 1943 [1]. Dorfman developed the idea in order to help the United States in combating outbreaks of syphilis in soldiers during World War II. One challenge to address is that group testing is less sensitive than individual testing, increasing the number of false negatives and sacrificing accuracy for speed. As we will see, however, this method would still have a significant impact on our testing capacity for COVID-19 and early detection of an outbreak.

Suppose that we would like to test N people in a population for a virus. Our strategy is to put k individuals in a group for $1 \leq k \leq N$, giving $\frac{N}{k}$ groups of k people. Combine the samples of these k individuals and test this combined sample. In the first round of testing, we administer $\frac{N}{k}$ tests. If a combined group tests positive, then we administer the test to each of the k individuals from that group.

Now suppose that the probability that an individual has the virus is given by p. Provided each administered test is independent of the others and our test is 100% accurate, we can model the situation using Bernoulli trials. Given a

group of k individuals, the probability that the group tests positive for the virus is then given by $1-(1-p)^k = 1-q^k$ where $p+q=1$.

Of courses, our main concern will be the expected value of the random variable given by X where X is the number of tests administered. We would like to minimize testing while also catching early outbreaks with high probability. First consider the situation where each group test is 100% accurate. For a given group of k individuals, if the group tests negative, we only need one test. If the group tests positive, we need $k+1$ tests. Since there are $\frac{N}{k}$ group of k individuals, by assigning appropriate weights, it follows that our desired expected value is given by

$$E(X) = \frac{N}{k}\left((k+1)\times(1-q^k)+1\times q^k\right) = \frac{N}{k}\left(k+1-kq^k\right) = N\left(1+\frac{1}{k}-q^k\right).$$

Since we are trying to minimize cost, we are trying to find the value of k that minimizes $E(X)$, for $1 \leq k \leq N$. Thus we have

$$\frac{dE}{dk} = N\left(-\frac{1}{k^2}-q^k\log q\right) = 0$$

which gives

$$\frac{1}{k^2} = q^k \log\left(\frac{1}{q}\right).$$

In order to approximate a solution for k then, we use available data to input $q = 1-p$. Whether or not a solution exists within our parameters $1 \leq k \leq N$ depends on the value of q. It is also of interest to note that the value of k that minimizes $E(X)$ is independent of N, the size of our population under consideration.

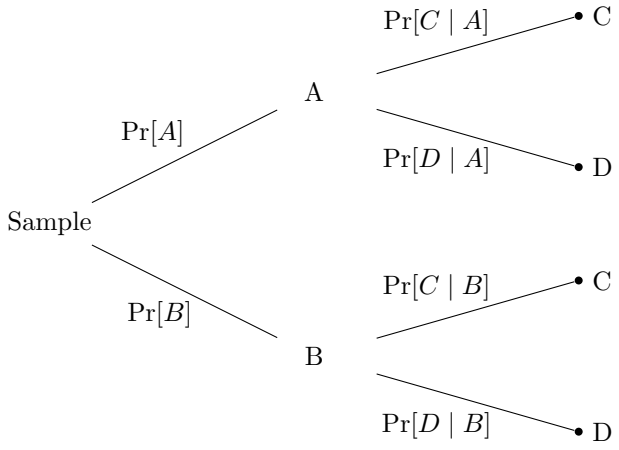
2 False Negatives and False Positives

In reality, such a group test is not 100% accurate. Preliminary methods suggest successful detection to groupings of up to 64 people [4]. The optimal size for k, for a broad range of p-values will be seen to be far below this number. Our concern with pooling samples together is that it may dilute the group sample and diminish the accuracy of the test. While this puts a damper on the method, one still may achieve significant reduction in cost (i.e. number of tests administered) while detecting outbreaks early. The derivation to minimize $E(X)$ above presumes that for a combined sample, if one individual is positive then the group tests positive and if all individuals are negative then the group tests negative. We aim to analyze how the existence of false positives and false negatives affect the outcome.

While false positives present a financial obstruction to assessing a given population's rate of infection, false negatives are a concern of greater gravity.

Certainly, if a perfect picture of where the virus resides at a given moment is the goal, false negative are an existential problem. However, if detecting an outbreak in its early stages, or keeping the spread of the virus under a certain percentage r, where r is small, is the goal, then group testing could prove worthwhile. Group testing will present a loss of accuracy in place of speed and cost reduction, a transaction that, under the right circumstances, including the current situation regarding COVID-19, is worth making.

Denote our events as follows: A is the event that the (group) sample is positive, B is the event that the sample is negative. C is the event that the sample tests positive and D is the event that the sample tests negative. We appeal to the probability tree below for calculations.



where $\Pr[A] = 1 - q^k$ and $\Pr[B] = q^k$. Thus we have

$$\Pr[A \cap D] = \Pr[A] \Pr[D \mid A] = (1 - q^k) \Pr[D \mid A]$$

and also

$$\Pr[B \cap C] = \Pr[B] \Pr[C \mid B] = q^k \Pr[C \mid B].$$

The probability that a given group of k individuals tests positive is thus

$$\begin{aligned} \Pr[C] &= \Pr[A] \Pr[C \mid A] + \Pr[B] \Pr[C \mid B] \\ &= (1 - q^k) \Pr[C \mid A] + q^k \Pr[C \mid B]. \end{aligned}$$

We are still interested in the expected value of the random variable X, where X is the number of tests administered. Adjusting appropriately, we have

$$\begin{aligned} E(X) &= \frac{N}{k} \left((k+1) \Pr[C] + 1 \times (1 - \Pr[C]) \right) \\ &= \frac{N}{k} ((k+1) \left[(1 - q^k) \Pr[C \mid A] + q^k \Pr[C \mid B] \right] \\ &+ 1 - (1 - q^k) \Pr[C \mid A] + q^k \Pr[C \mid B]). \end{aligned}$$

3 Applying Group Testing to COVID-19

We utilize recent data on COVID-19 parameters to make some estimations about how our above equations would look in practice. In particular, as the author currently resides in the state of Indiana, we will use data from Indiana. The reader may look up the appropriate values for replacement as desired. Recent estimates in Indiana suggest the value of $p = 0.028$ giving $q = 0.972$ [2]. For 1-to-1 testing of COVID-19, false positives appear to be exceptionally rare [3]. For a group test, such a false positive would be even more rare. With this in mind, we will approximate $\Pr[C \mid B]$ to be zero. Thus we presume

$$\Pr[C] = (1 - q^k)\Pr[C \mid A] = (1 - q^k)(1 - \Pr[D \mid A]).$$

Thus we have

$$\begin{aligned} E(X) &= \frac{N}{k}((k+1)\left[(1-q^k)\Pr[C \mid A] + q^k \Pr[C \mid B]\right] \\ &\quad + 1 - (1-q^k)\Pr[C \mid A] + q^k \Pr[C \mid B]) \\ &= \frac{N}{k}\left((k+1)(1-q^k)\Pr[C \mid A] + 1 - (1-q^k)\Pr[C \mid A]\right) \\ &= N\left(\frac{1}{k} + (1-q^k)\Pr[C \mid A]\right). \end{aligned}$$

To minimize our expected number of tests, we then solve

$$0 = \frac{dE}{dk} = N\left(-\frac{1}{k^2} - \Pr[C \mid A] q^k \log q\right)$$

or more simply

$$\frac{1}{k^2} = \Pr[C \mid A] q^k \log \frac{1}{q}.$$

This is the same equation we had in our introduction with the added factor of $\Pr[C \mid A]$. Of course, for a completely accurate test, we recover our original equation. Plugging in our estimation for q we obtain

$$\frac{1}{k^2} = (1 - \Pr[D \mid A])(0.972)^k \log\left(\frac{1}{0.972}\right).$$

For proof of concept, we select a false negative rate that is relatively large, say $\Pr[D \mid A] = 0.2$ so that $\Pr[C \mid A] = 1 - \Pr[D \mid A] = 0.8$. We then approximate the above equation to find our value minimizes at $k = 7$. Thus, if a college campus, call it CC, contained $N = 7,000$ faculty and staff, the expected number of administered tests would be minimized by grouping students into sets of 7 from which we would expect

$$7,000\left(\frac{1}{7} + 0.8(1 - 0.972^7)\right) = 2010$$

administered tests. Thus the cost would be only $\frac{2010}{7000} = 0.287$ or 28.7% the cost of testing each student, a significant reduction.

It is not hard to see that as $\Pr[D \mid A]$, the optimal value for k also increases. The problem is that this derivation is concerned with minimizing costs, not being accurate. We next ask ourselves how likely, under these conditions, it would be to miss an outbreak. An outbreak is when there is a sudden increase in the number of positive cases in a small geographical area. To demonstrate unlikeness of missing an outbreak under these conditions, we explicitly compute an example.

Let Y be the random variable denoting the number cases in which the group is positive but tests negative. Next recall that $\Pr[A \cap D] = (1 - q^k) \Pr[D \mid A]$. Suppose it is known that prior to an outbreak, 2.8% of the population of CC were estimated to be positive and we ask the probability that 4% is now positive and our group testing missing this fact. Let's ask, say, the likelihood of having 1.2% in false negatives. This would be 84 out of 7000. Given a grouping of $k = 7$, we assume that each group does not have more than one positive case. Otherwise the probability of a false negative would, in fact, be lower. We find, under the conditions of CC, we have $t = 0.2(1 - 0.972^7)$, and

$$\Pr[Y \geq 84] = \sum_{n=84}^{1000} \binom{\frac{7000}{7}}{n} t^n (1-t)^{\frac{7000}{7} - n} = 2.1 \times 10^{-12}$$

which is a negligibly small value. For situations larger than 4%, the likelihood of catching an outbreak is better and we find that our group testing is suitable. In addition to this, repeated group testing, such as once a week, would increase the chances of catching an outbreak even more as the loss of accuracy would be compensated by the frequency of testing.

References

[1] R. Dorfman, *The Detection of Defective Members of Large Populations*, The Annal of Mathematical Statistics, **14** (4): 436-440, 1943.

[2] Retrieved from `https://news.iu.edu/stories/2020/05/iupui/releases/13-preliminary-findings-impact-covid-19-indiana-\coronavirus.html`

[3] Retrieved from `https://www.unionleader.com/news/health/coronavirus/state-official-false-positive-covid-19-tests-very\-rare-not-so-with-false-negatives/article_369216dc-3a09-5139-8a29-0520bc3c9d63.html`

[4] Retrieved from `https://www.youtube.com/watch?v=vxs11ryS9Dg&feature=emb_logo`