1 General introduction

Late 2019 marked the beginning of the spread of SARS-Covid19 where it first appeared Wuhan China then quickly spread to the rest of the world. Because of the nature of the disease its spread was rapid. Many countries suffered from it, the health systems and facilities have been overwhelmed and failed to meet the demand. To control its transmission governments imposed many measures but the rapid detection of the infectious cases is crucial to contain the disease. Therefore, governments and health organizations conducted massive testing on the population. With this comes another problem that is, testing of this scale is very expensive and needs big capacity.

Group testing or pool testing has been used in many countries. This was first studied by Robert Dorfman in the United states during the second world war, where a large number of new recruits of the army were tested for Syphilis [citation].

The samples to be tested are taken from a group (pool) rather than individuals. The population is divided into groups and then a sample representing a mixture of each individual sample in the group is tested. If a test result of a particular mixture is negative then all the individuals in the group are free of the disease. If the test comes out positive, then at least one of individual the group is infected. Several tests of such kind can be used to determine the infected individuals of the population in contrast to the traditional way where each individual is tested alone. With arranging the procedure carefully as precise tests, good selection of groups, this would reduce the number of tests dramatically.

In general, suppose that we have n number of individuals to be tested. The traditional way is to conduct n number of tests. If the number of infected is large then this would be reasonable. By the time of Dorfman's work, the U.S army was testing the new recruits for Syphilis, and it was rare among the population and he started studying pool testing[refer to the information theory].

The problem can be described as, given a number of items (persons) n and number of defectives (infected) d, then how many tests t are required to discover the number of defected (infected) among the n items (persons) and what is feasible way to achieve this [also to information theory].

This problem can be classified depending on the point of view that is taken into consideration.

2 Mathematical background

The problem is to reconstruct the missing information (detection) from the measured data. The basic assumption is that the information is obtained and processed following a linear model, that is the problem is reduced to solving a system of linear equations. Suppose that the results from the laboratory are referred to as $y \in \mathbb{R}^n$, which corresponds to $x \in \mathbb{R}^n$. x represents the information about the detection of the infected.

The problem is formalized as follows:

$$Ax = y \tag{1}$$

The matrix $A \in \mathbb{R}^{m \times n}$ is called the design matrix, which represents the testing process, m is the number of tests and n is the number of people tested [reference:the mathematical book].

The aim is to recover the vector $x \in \mathbb{R}^n$ by solving Eq.(1) for x. Ideally, $m = n$ hence the system is easily solved but it is common to have $m < n$ since the problem states that. Linear algebra tells us that this system is undetermined and it has infinitely many solutions. This renders the problem impossible to solve, *i.e.* recover the exact vector x. However, under certain assumptions it is possible to have a solution of the problem, and there are efficient algorithms to recover x exactly. The assumption is the referred to as the **Sparsity**.

Sparsity: a vector is said to be sparse if most of its components are zeros. Formally, a vector $x \in \mathbb{R}^n$ is called $s - sparse$ if at most s of components are non-zeros. Linking this to our work is, since the vector x in 1 represents the individuals to be test among which some of them are infected, with knowledge of the infectious rate and it assumed to be small, then with zero represents being free of the disease, then it is appealing to consider that our vector is sparse [the mathematical book].

In general, looking at the problem of reconstructing x, one would address two main questions:

- What is the best matrix $A^{m \times n}$ that one uses to obtain the best results?
- What is the efficient algorithm that one uses to reconstruct x from $Ax = y$?

3 Adaptive and non-adaptive Group testings

3.1 Adaptive (sequential) tests

In Adaptive (sequential) algorithms, the tests are conducted sequentially one by one. The results from one test are recorded and used in the next test [reference ku book]. For instance, imagine a situation of testing 100 individuals for Covid19. At the beginning two groups are formed and one assumes one of the groups is free of the virus. thus for the second round of testing the knowledge from the first test is used as that, all the individuals in the first group are discarded and the second group is divided into smaller distinct groups.

3.2 Non-adaptive tests

In the non-adaptive algorithms, the groups to be tested are determined in advance, the tests are conducted in parallel and an individual can be involved in more than one test. It is intuitive that non-adaptive tests are time saving, and in our case, the covid19, where the tests must be conducted simultaneously, this is advantageous.

However, in general the adaptive tests are faster than non-adaptive because it transfers information from the previous tests to next tests. Considering the advantages from both types it is appealing to try to treat one test like the other. Non-adaptive tests can be treated as adaptive by staging the tests. Here the tests are divided into stages and information is transferred from one stage to the next, but within the stage information is not shared[ku book].

Although the two questions raised above are connected since the design matrix is part of the algorithm to be used, it is often helpful to try to answer them separately. Next we introduce common terminologies of the groups testing models that determine our definition of the model being successful i.e. achieving the best results with a minimum number of tests.

1. Binary and non-binary: In the standard group testing model, the design matrix, the defective vector and the outcomes vector are all in binary form. Where 0 in the design matrix represents not being tested and 1 otherwise, while in the defectives and the outcomes vectors, 0 for being free of the virus and 1 otherwise. Other models consider non-binary representation. In work here we will use binary representation[reference]. The use of the binary setting is beneficial whereas it allows to use the literature of compressive sensing problem.

2. Noise: Noiseless are tests where there is no error of any kind, that is the recovery is done complete i.e. the test procedures are done perfect. while in noisy tests we expect errors to happen either from the model or other reasons for example human errors. Two common model errors are the false positive (specificity) or false negative (sensitivity), where the former the negative outcome is labeled positive by the algorithm and later a positive outcome is labeled negative. In the realistic models small error probability is allowed, where the set of defective individuals is taken to be the defectives with high probability[reference].

3. In our work here we consider a situation where the number of infected is known in advance through the knowledge of the infectious rate. The problem is combinatorial as the set of defectives (infected) is uniformly random among the sets of its size.

Our work will focus on the combinatorial non-adaptive tests and its implementation by building in-silico simulations of a non-adaptive test.

4 Key words

1. **Defective set**: We refer to the number of elements to be tested as n. We write $D = \{1, 2, ..., n\}$ for the set of defectives where each number item in the list is a label given to each individual and $d = |D|$ is the number of defectives.

   We write $x_i = 1$ for the defective (infected) item $i \in D$ and $x_i = 0$ for non defective (infected) item$i \notin D$.

2. **Tests**: we refer the number of tests performed or to be performed by m, and label the tests $\{1, 2, 3..., m\}$. The tests represent the rows of the matrix A where the element a_{ij} is element j of n and tested in the i^{th} test. An Important consideration is that each item is included in each test independently with fixed probability. The matrices we will use are sparse matrices. Although there is no particular criterion to to judge exactly if a matrix is sparse or not it is common that to set the number of non zero elements to be equal to the number of rows or columns of the matrix. Note that through out this text we will refer to this matrix by A or if else it will be stated. s

3. **Standard noiseless group testing model** for known m, n and a test matrix $A^{m\times n}$ this model is described by its outcomes as follows:

$$y_j = \begin{cases} 1 \quad if \quad \exists i \in D \quad \text{with} \quad x_{ij} = 1 \\ 0 \quad if \quad \exists i \in D \quad with \quad x_{ij} = 0 \end{cases}$$

where $y_j \in \{0,1\}^m$ is the test outcome.

4. **Detection algorithm**: It is a function (map) that maps an element of the set $\{0,1\}^{m\times n} \times \{0,1\}^m$ to the subsets of $\{1,2,...,n\}$. Under an ideal recovery the output of the function above is the set of defectives (infected) [ref here].