Team Control Number

1234

Problem Chosen

A

2021
HiMCM
Summary Sheet

Abstract

One of the most important issues in mitigating the COVID-19 pandemic is testing. In some countries, regions or cities, the problem of insufficient detection capabilities is often raised. In order to use resources more effectively, we consider the use of **group test methods**.

For problem 1, we first established a **two-stage model, a three-stage model, and a multi-stage model** without considering the detection error, and obtained the best group size under different positive rates p. Then we considered the case of detection errors, assuming the error rate is α, we modeled the **false positive rate and false negative rate**, and obtained the quantitative relationship between them and the group size. In the case of comprehensively considering the three variables, we established a **constrained optimization problem** and the **penalty function method** is used to get the best group size.

For problem 2, compared with ordinary patients, asymptomatic infections cannot be assisted by means of **"clinical diagnosis"**, which can effectively reduce the false negative rate fnr, and we find the false negative rate of group detection is higher than that of non-group detection, so We can only hope that the false-negative rate of grouped nucleic acid testing is a little lower. We have proposed a **"multiple testing at the mixed inspection stage"** scheme, and compared with the original scheme, we found that it can indeed greatly reduce the false-negative rate of group test.

For problem 3, we need to consider proposing a new group test scheme to further improve the detection efficiency. The solution we propose is **an improved two-stage group test**, and there will be **duplication** between different groups. Specifically, the people to be tested are arranged in a square, each row and column of the square are taken out for testing separately, and finally the confirmed patient is determined by the intersection of the row and the column. Taking Hong Kong as an example, we show through numerical experiments that when the positive rate p is small, this improved two-stage grouping detection scheme (requiring 964,922 tests) is lower than the original scheme (requiring 1,467,682 tests), which shows this is **more efficient**.

Since the positive rate p and the detection error rate α are estimated, the **uncertainty is high**, so we conducted a **sensitivity analysis** on the positive rate p and the detection error rate α. It is found that the sensitivity of the total number of detections m to the positive rate p and the detection error rate α is about 0.5 and 0.1, much small than 1, so our model is **very stable**. We have also analyzed the advantages and disadvantages of the model. Our model is considered **comprehensive**, and we have drawn **a lot of diagrams** to show the relationship between various variables. Finally, we wrote a letter to the Hong Kong government organization, detailing our research results and recommendations for them.

Keywords: COVID-19; group test; false positive; false negative; penalty function

Contents

1 Introduction

1.1 Background

Coronavirus, which is also known as COVID-19, is pneumonia caused by novel coronavirus infection. It is a very infectious virus which therefore became a pandemic admitted by W.H.O. (World Health Organization) and now one of the biggest problem is the detection.

There are two kinds of results, one is positive, one is negative. A positive indicates infection, while a negative indicates no infection. Usually the medical staff will take a sample from each person and after the test, they will issue a notice to each person informing them of the results. Detection is a trouble to lots of countries, areas and cities usually due to the lack of resources. For example, a typical example could be Hong Kong, during the third wave of COVID-19, experts recommended testing 20,000 times a day, but only 4,000 to 5,000 times a day were supported at the time. In order to make sure that the daily test population reaches the target, grouping is extremely important.

So why could grouping reduce the number of tests? Let's suppose among 100 people who needs to be tested, there is one positive and 99 negatives, our goal is to find out who is the positive one. If we use the traditional method to test them one-by-one, 100 times are required to test them all. But if we divide 100 people equally into five groups, then we can just pick the group that is positive and test each person in that group to determine who is positive.The whole process requires only 25 tests, far less than 100. If this scheme can be put into practice, a lot of resources can be saved.

Here we consider schemes with multiple stages, where stages refer to how many times they are grouped. If no positive samples are detected in a group at a certain stage, the entire group is confirmed as negative and need not proceed to the next stage. The final stage of the detection should be one person per group, that means everyone in those groups that show up positive in the penultimate phase needs to be tested. Therefore to sum up, the test is only over when every positive sample has been tested and everyone knows their results.

In order to ensure the feasibility of the protocol, we also need to consider the test error, usually false positive and false negative error. A false negative is when a positive person is tested negative, and a negative tested positive is a false positive. The false negative is more harmful because if you put a positive person back into society, that person will infect more people, and there will be more people to be tested. Therefore, a good scheme also needs to prevent the occurrence of false test.

1.2 Problem Restatement

1. Due to resource constraints, it is difficult for the government to do nucleic acid testing for every resident. At this time, grouping people is particularly important, because it can effectively reduce the number of tests. Then how to minimize the number of nucleic acid tests while everyone knows their results? Sometimes a test can turn a positive person into a negative, called a false negative, or a false positive the other way around.So what will happen if the tests contain errors like false positives and false negatives?

2. COVID-19, unlike other kind of diseases, contains asymptomatic infected people who in the first few days of infection are normal and undetectable, which could cause serious effect to people around them. So how to solve the problem that asymptomatic infected persons cannot be detected?

3. Besides schemes with multiple stages, there could also be single stage schemes. Even when grouping people, it is possible to put one people into two or even more groups to make sure the results are accurate. Therefore a better scheme is needed for grouping.

4. Write a letter to Hong Kong government's department of health in order to provide them with some advice.

1.3 Our Work

To solve this problem, we first considered the situation where there is no error in the detection, established two-stage, three-stage and multi-stage models with probability and expectation methods, and then considered the situation of detection errors, and established a model with error rate. Then, for the asymptomatic infected population, a method of multiple tests in the mixed test stage was proposed to reduce the false negative rate. Finally, a new group detection scheme was proposed, and Hong Kong was used as an example to explain why this scheme is better under certain settings. Finally, a letter was written to the Hong Kong government to introduce our findings and conclusions.

2 Assumptions and Justification

- **Assumption 1:** Assume that the positive rate p in the population can be obtained more accurately through a small-scale test. The positive rate in the population is a very critical variable in our model and plays a key role in the size of the group. The positive rate p is generally obtained by random sampling from the population. As long as the sampled population and sample size are appropriate, the positive rate in the population can generally be obtained more accurately. However, there may be differences in the positive rate of p in different regions, and it is necessary to design a sampling plan for the region.

- **Assumption 2:** Assume that the positive rate p in the population cannot be too large. Our group detection scheme is based on the fact that the positive rate p in the population cannot be too large, because we will analyze below that once the positive rate p is too large, group detection will not be significant compared to non-group detection Advantage. This assumption is also reasonable, because at this stage the infection rate of the new coronavirus is not very high in most parts of the world, especially the positive rate in China is very low. If the positive rate of p is very large, it is likely that we humans have entered a state of symbiosis with the virus, which has not happened yet at this stage.

- **Assumption 3:** Assume that the error rate of nucleic acid detection α is a constant. Relevant literature indicates that false positives and false negatives in nucleic acid testing are caused by detection errors, and they may not be the same. However, it is very difficult to accurately obtain the error rate of the two. In fact, the difference between them is not significant. Therefore, for the convenience of modeling, it is reasonable to regard it as a unified error rate α.

- **Assumption 4:** Assume that the detection error rate of asymptomatic infected persons is the same as that of ordinary patients when undergoing nucleic acid testing, but it cannot be detected by clinical diagnosis. Relevant literature shows that asymptomatic infections can be well detected by nucleic acid testing, but there are no clinical symptoms in the short term. Therefore, in nucleic acid testing, we treat asymptomatic infections and ordinary patients uniformly, without distinction, they are only different in clinical diagnosis.

3 Variables

Table 3.1: Variables Table

Variables	Description
n	Total number of samples
m_1	Average number of detections for a group
m	Total number of testing
k	Group size for group test
p	Positive rate in the population
q	Negative rate in the population
α	Nucleic acid detection error rate
fpr	False positive rate
fnr	False negative rate
$E(P)$	Expected number of positive patients
$E(N)$	Expected number of negative patients

4 Modeling

In order to establish a model aimed at reducing the total number of tests, we decided to first establish a two stage model and a three stage model with probability and expectation, so as to clarify the relationship between each variable, and then add the occurrence of false positive and false negative to build this complex model.

4.1 Basic model

This section contains two stage model and three stage model used to lay the ground work for the model with test error rate.

The first thing we consider is a two-stage model. We always start with simple models, which will help us solve problems. First of all, in the first step of the model, we need to determine the positive rate p in the population, which is very critical to our problem. Because we know that group detection is not always a good method, it depends on the positive rate in the population p, we will do a detailed analysis below. The method to get the positive rate in the population is to conduct a small-scale test. Suppose that n_1 individuals are selected from the population of n for nucleic acid testing, and it is found that there are n_2 individuals who have been tested positive for nucleic acid, then the population is positive The rate can be estimated as:

$$p \approx \frac{n_2}{n_1}$$

We should note that the selection of n_1 has an important impact on the estimation of the positive rate p. Under normal circumstances, choosing a larger n_1 can better estimate the positive rate in the population p, but this will require the support of manpower and material resources; choosing a smaller n_1 for the positive rate in the population p will be less accurate, but at the same time it saves manpower and material resources. In short, n_1 should choose a moderate value, which will be related to actual problems. For example, for Shanghai with a population of 20 million, you can choose $n_1 = 20000$ to estimate the positive rate p. Anyway, suppose we have got a more accurate positive rate p.

4.1.1 Two stages model

If we divide the model into two stages, the model is shown below.

When all the mixed samples are negative and no further testing is required, the probability is as follows:

$$p_1 = (1-p)^k \tag{4.1}$$

On the contrary, when the mixed samples are positive and no further test is required, the probability is as follows:

$$p_2 = 1-(1-p)^k \tag{4.2}$$

Therefore the average number of tests required for a mixed sample is shown below:

$$m_1 = p_1 \times 1 + p_2 \times (k+1) = q^k + (1-q^k) \times (k+1) = k+1-kq^k, \tag{4.3}$$

in which

$$q = 1-p \tag{4.4}$$

So from the above results, we can get the total number of detection, which is shown in the formula below:

$$m = m_1 \times \frac{n}{k} = n + \frac{n}{k} - n \times q^k = n[1 + \frac{1}{k} - (1-p)^k] \tag{4.5}$$

For different p, we can draw the following picture: the total number m vs the group size k.

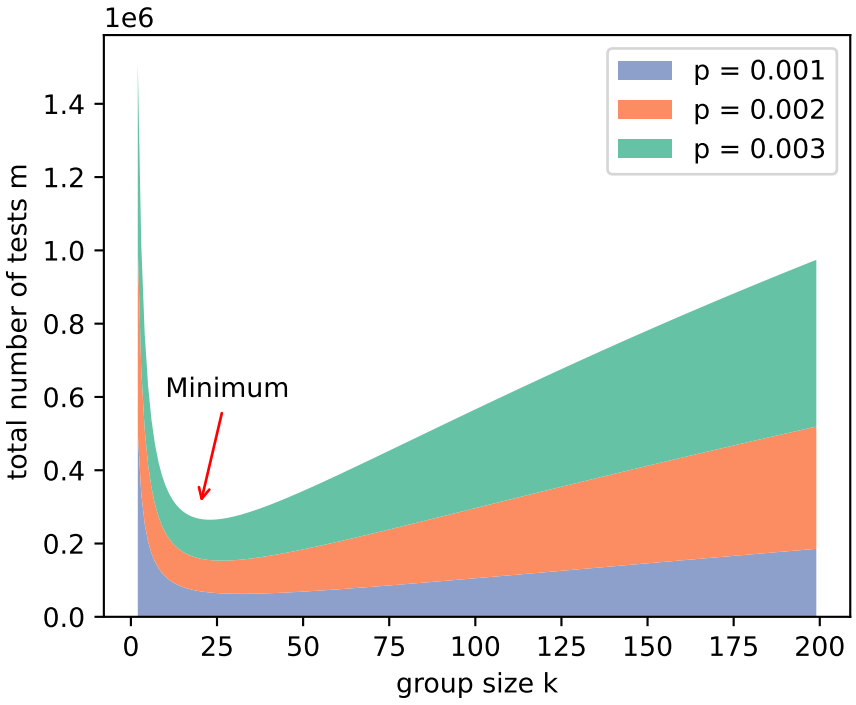


Figure 4.1: the total number m vs the group size k

From this picture, we can see that the function is convex. For each fixed p, there will be a unique k, which makes the total number of detections the least. For example, for $p = 0.003$, the optimal group size is about 20.

In order to obtain the best relationship between the group size k and the positive rate in the population, we need to require the derivative and set the derivative to 0

$$m' = \frac{dm}{dk} = -\frac{1}{k^2} - q^k \ln q = 0 \tag{4.6}$$

But we cannot get the relationship between k and p from it, because it is difficult for us to figure out k.

Luckily, if the possibility of testing positive is small, then we will have an estimate:

$$(1-p)^k \approx 1 - kp \tag{4.7}$$

We use the following experiment to prove why this is right.

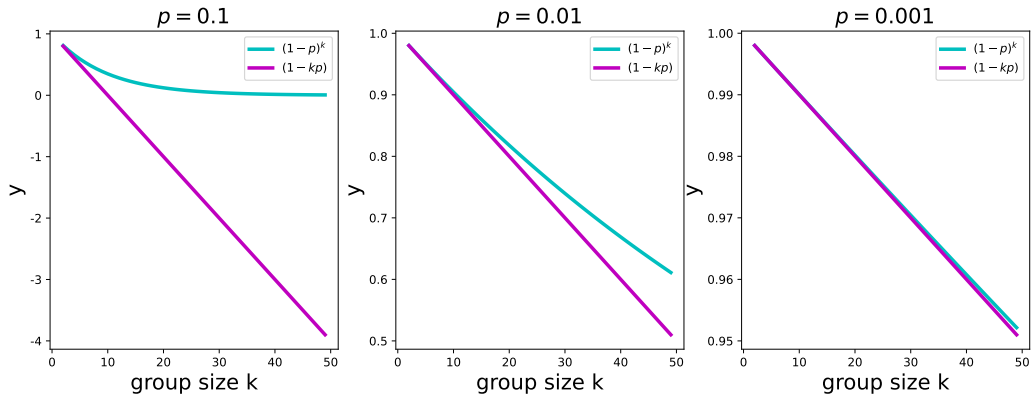


Figure 4.2: Approximate relationship

Purple is the line $1 - kp$, and cyan is the function $(1-p)^k$. As you can see, as p decreases, these two functions are quite close. Therefore, we can rewrite the expression of the total number of tests through this approximate relationship.

$$m = n(\frac{1}{k} + kp) \tag{4.8}$$

We recalculate the derivative of m with respect to k:

$$m' = \frac{dm}{dk} = n(-\frac{1}{k^2} + p) = 0 \implies k = \sqrt{\frac{1}{p}} \tag{4.9}$$

We see that k and p are similar to inversely proportional relationships, and here's a root number. This is very much in line with our actual lives, and we intuitively know that when the positive rate in the population is very low, it is possible to improve testing efficiency by increasing the group size by k, which is consistent with our model results.

We've drawn the relationship between the optimal group size k and the positive rate p in the population, as shown in the following image. As you can see, the function image is similar to the inverse function image.

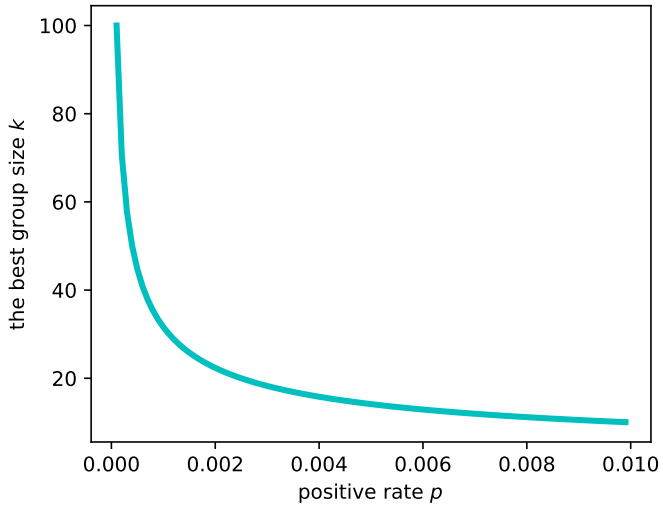


Figure 4.3: the optimal group size k vs the positive rate p

Next, we naturally have a question, **does group detection always improve detection efficiency?**

As can be seen from the figure above, when the positive rate in the population is p is very large, the optimal group size may be 1.So here's an example of how group testing doesn't always improve test efficiency when the positive rate in a population is p large.

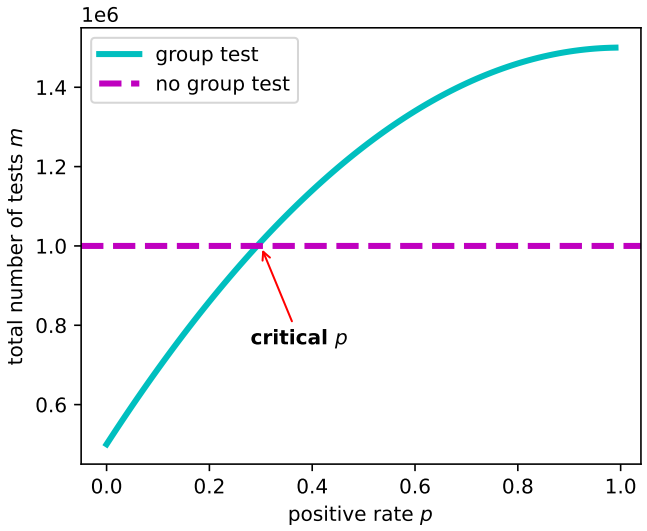


Figure 4.4: Fixed $k = 2$, group test vs no group test

As shown in the figure above, dashed lines are not grouped for detection, there is a critical p value that makes group detection less efficient than non-group detection, where we fixed $k = 2$.

4.1.2 Three stages model

When there are too many people to test, the two-stage model is not enough so we're going to introduce a three-stage model.

If the first mixed detection is negative, then no more detection is required, the probability is listed below

$$p_1 = (1-p)^{k_1} \tag{4.10}$$

If the first mixed detection shows positive, then additional detection is needed.

When the second detection appears negative, there will be no need for additional detection.Conditional probabilities are shown below. Denote that:
A: the first detection is positive
B: the second detection is negative

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{q^{k_2}(1-q^{k_1-k_2})}{1-q^{k_1}} \tag{4.11}$$

If the second detection is positive, each sample needs to be tested. Bellow is the conditional probability for this situation.

$$P' = 1 - P(B|A) = 1 - \frac{P(AB)}{P(A)} = 1 - \frac{q^{k_2}(1-q^{k_1-k_2})}{1-q^{k_1}} \tag{4.12}$$

The average detection number of k_2 samples is the **expectation**:

$$m_1 = 2 \times \frac{q^{k_2}(1-q^{k_1-k_2})}{1-q^{k_1}} + (2+k_2) \times (1 - \frac{q^{k_2}(1-q^{k_1-k_2})}{1-q^{k_1}}) \tag{4.13}$$

Therefore the total number of detection should be the equation below.

$$m = [m_1 \frac{k_1}{k_2} + (1-p)^{k_1} \times 1] \frac{n}{k_1} \tag{4.14}$$

Now we're looking for the minimum value of the function m, with the arguments being k_1, and k_2 being integers.

Now the function has two dimensions but we can convert it into a variable, namely taking the $k_2 = \frac{1}{10}k_1$, so that there is only one variable k_1. We can draw a function image of k_1 and m, and finding out the smallest function value and the corresponding k_1.

Corresponding to different p, we can draw the function graph as follows. It can be seen that the function is still convex, which is basically the same as the two-stage analysis. There is a unique k_1 that minimizes the total number of detections.

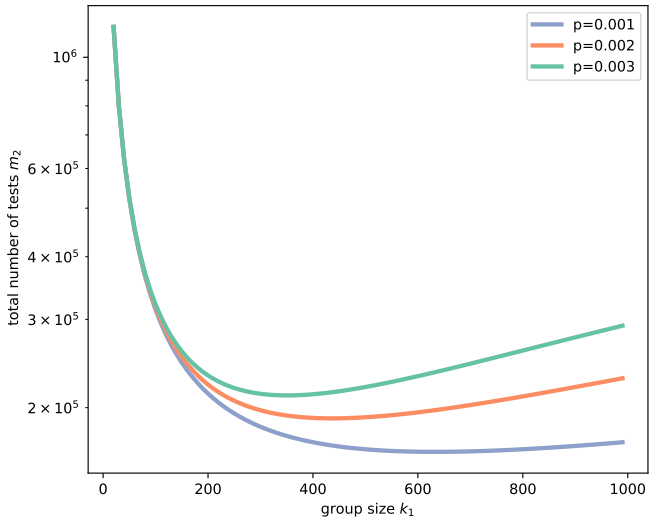


Figure 4.5: m vs k for 3 stages

In addition, we can also solve it directly as a two-dimensional function problem. At this time, the function image will be three-dimensional, and we have drawn the following schematic diagram through the three-dimensional drawing function of python.

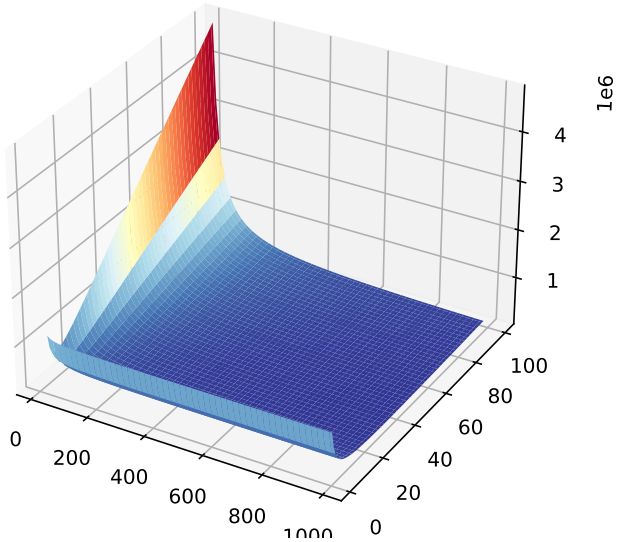


Figure 4.6: m vs k_1 and k_2. Red represents a larger value, and blue represents a smaller value.

From the three-dimensional image, we can clearly see that the function value at the edge is very large, which corresponds to a large number of detection times. The minimum value of the function is difficult to obtain through image observation.

By taking the minimum value of the array, we find that when k_1 takes 340 and k_2 takes 97, the total number of detections is the least.

Below we draw another bubble chart. This is obtained by randomly picking points within a certain interval. The size of the bubble represents the size of the function value, and the color of the bubble has no special meaning. Similarly, we see that the bubbles at the edges are larger, indicating that the total number of detections is larger. At $k_1 = 340$, $k_2 = 97$, the bubbles are smaller, which corresponds to a smaller number of detections.

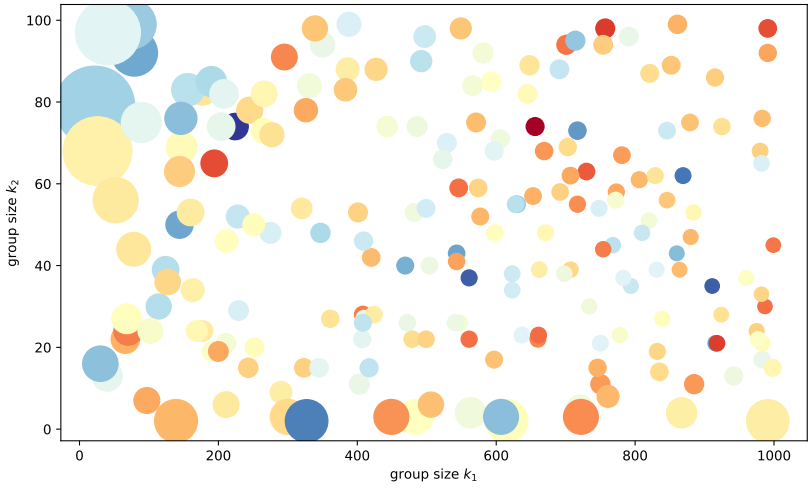


Figure 4.7: The bubble chart. The size of the bubble represents the size of the function value.

4.1.3 Multi-stage model

If the number of stages continues to increase, it will become extremely difficult to analyze the problem from a probability perspective, because each stage will be divided into two results: negative and positive, so there will be 2^M kinds of results in the M stage. Conditional probability calculation is very difficult. We analyze the problem of multi-stage detection from another perspective. We consider the worst case and make an upper bound analysis of the total number of detection.

Similarly, we still start with two stages. In the first stage, you need to check n/k times. Those with negative results in the first stage do not need to participate in the second stage of testing. We don't know how many groups of test results will be positive in the first stage, but we know that this will never exceed np, because there are only np patients in the population, and the worst case is that all np patients be divided into different groups. So in the worst case, the number of detections is

$$m = \frac{n}{k} + npk \tag{4.15}$$

We magically discovered that this is exactly the same as the result (4.8) after we have done the approximation before.

We know that this function is shaped like a "check mark", so it has a name called "check mark function", as shown in the figure below, there is a unique minimum point k.

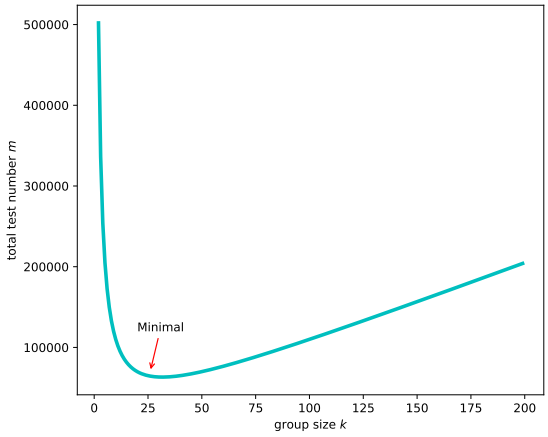


Figure 4.8: m vs k for a 2-stage upper bound analysis

Now we consider the problem of the $M+1$ stage. This seems relatively complicated, but with the upper bound analysis above, we can get the following results with the same idea:

$$m = \frac{n}{k_1} + \frac{npk_1}{k_2} + \cdots + \frac{npk_{M-1}}{k_M} + npk_M \tag{4.16}$$

Let's explain a little bit about this formula. Except for the first stage, we don't know the number of positive groups in every subsequent stage, but we know that this will never exceed np, so we all use np as an upper bound.

Now a difficult question arises. We need M variables $k_1, k_2, \cdots, k_M$, it is very difficult to solve them directly as a M dimensional problem. Even if we can solve the result, it is not practical in real life. Therefore, using a similar approach as in the previous section, we artificially add a restriction $k_{i+1} = 1/2k_i$ so that there is only one variable k_1, and we can draw the corresponding to different stages number M, the relationship graph between the total number of detections m and k_1.

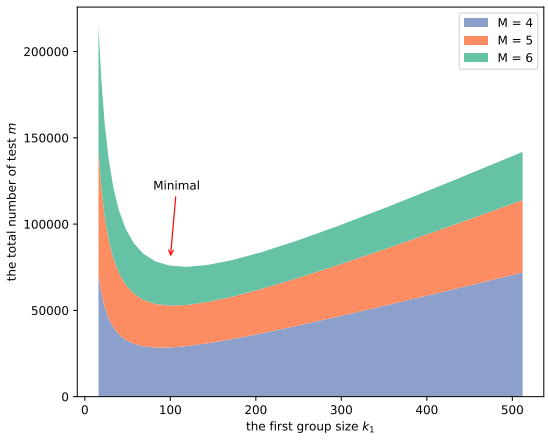


Figure 4.9: m vs k for a multi-stage upper bound analysis

We see that an increase in the number of stages does not mean a better solution, because the total number of detections has increased. Taking into account the actual operation, multi-stage grouping is difficult to perform, because it may confuse the samples, so our main model is still to consider two-stage and three-stage.

4.2 Model with test error

What we all consider above is that the test results are correct, but in reality, the results of nucleic acid tests are not all correct. It is possible that a person's own nucleic acid is negative, but what he gets is a positive test report, which is called "false positive". It is also possible that a person is positive, but got a negative test report, which is called "false negative". Now we consider the situations in which detection errors may occur. For simplicity, we can assume that the detection error rate is a constant α. This can be obtained through small-scale population testing or evaluated by developers.

		Standard positive	negative	Total
Result	positive	a	b	$a + b$
	negative	c	d	$c + d$
Total		$a + c$	$b + d$	n

The table above shows the different results of detection, which contains not only positive and negative, but also false positive and false negative. Standard means the actual situation of a person and result means the result of detection, which may sometimes be wrong. According to the definition of false positive, the number of false positive should be b which means a negative standard corresponds to a positive result, in contrast, c means the number of false negative.

If we want to calculate the false positive rate (fpr) for everyone, we divide the number of standard negatives by the number of false positives, The false positive rate is

$$fpr = \frac{\mathrm{FP}}{\mathrm{FN} + \mathrm{TN}} \tag{4.17}$$

where FP is the number of false positives, TN is the number of true negatives. If we plug in the data from the list above in to this equation, then the rate of false positive and false negative is shown bellow,

$$fpr = \frac{b}{b + d} \tag{4.18}$$

$$fnr = \frac{c}{a + c} \tag{4.19}$$

If we know that there are n individuals with positive rate of p, where the detection error rate is α, then the false positive and false negative rates in the case of non-grouping are as follows

$$fpr = \frac{nq\alpha}{nq} = \alpha \tag{4.20}$$

$$fnr = \frac{np\alpha}{np} = \alpha \tag{4.21}$$

4.2.1 The total number of tests

Below we still consider the two-stage model. Suppose that k samples are mixed together, taking into account the required number of tests and the false-positive rate and false-negative rate. According to the result get from chapter 4.1.1, the possibility of negative mixed sample is

$$p_1 = (1-p)^k = q^k \tag{4.22}$$

As the result can be divided into negative report and positive report, we will see them separately.

1. If the report shows negative, two situations are provided. In the first case, they are all negative, indicating that the test results are correct. The probability is shown below,

   $$p_{11} = (1-p)^k(1-\alpha) \tag{4.23}$$

   In the second case, there was actually a positive in k individuals, but the test report was negative, so the test was wrong. The possibility follows,

   $$p_{12} = (1-q^k)\alpha, \tag{4.24}$$

   where $q = 1-p$. Therefore the total probability of negative result should be the sum of these two,

   $$p_1 = p_{11} + p_{12} = q^k(1-\alpha) + (1-q^k)\alpha \tag{4.25}$$

2. When the report shows positive, people are actually all negative in the first situation, however as the report shows positive, the detection must be wrong. The probability follows

   $$p_{21} = (1-p)^k\alpha \tag{4.26}$$

   For the second situation, because actually there are positives in k people, therefore the detection is correct, the probability is bellow,

   $$p_{22} = (1-q^k)(1-\alpha) \tag{4.27}$$

   So the total probability shall be,

   $$p_2 = p_{21} + p_{22} = q^k\alpha + (1-q^k)(1-\alpha) \tag{4.28}$$

From what has been discussed above, the expected number of times the mixed samples need to be tested is

$$m_1 = p_1 \times 1 + p_2 \times (k+1) = k + 1 - kq^k - k\alpha + 2kq^k\alpha \tag{4.29}$$

As we said above, there are n people in total, so the total numbers of times shall be

$$m = m_1 \times \frac{n}{k} = n + \frac{n}{k} - nq^k - n\alpha + 2nq^k\alpha = n(1 + \frac{1}{k} - q^k - \alpha + 2q^k\alpha) \tag{4.30}$$

Here we can see that when $\alpha = 0$ happens to be the result of our analysis above.

Below we give a graph corresponding to different detection error rates α, the total number of detections changes with the size of the group.

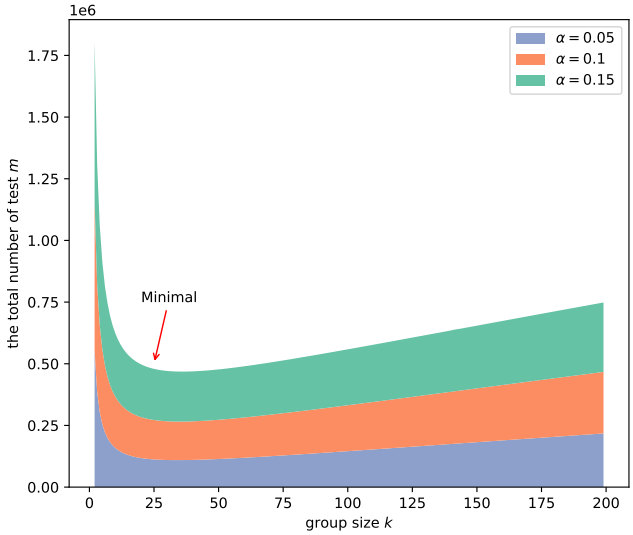


Figure 4.10: m vs k for a different test error α

We can see that as the detection error rate α increases, the minimum number of detections required also increases.

Similarly, we can also consider the approximate relationship. When the positive rate p is very small, there is

$$(1-p)^k \approx 1 - kp$$

So the above formula becomes

$$m = n(\alpha + \frac{1}{k} + kp - 2\alpha kp)$$

This is also a checkmark function, but the minimum value has changed. Derivation of k directly, you can find the analytical expression of k.

$$m' = \frac{dm}{dk} = -\frac{1}{k^2} + p - 2\alpha p = 0 \implies k = \sqrt{\frac{1}{p(1-2\alpha)}}$$

The optimal group size is an increasing function of the positive rate p and the detection error rate α in the population.

4.2.2 False positive rate and false negative rate

Next, consider the false positive rate and false negative rate of group detection. First of all, we first solve the following problem:

Problem Suppose the positive rate in the population is p, and the group size is k. It is known that there are positive patients in these k individuals. Let the random variable P represent the number of positive patients in the mixed test sample, and find the expectation of the random variable P. Here we use the conditional probability to calculate a little (the calculation process is listed in the appendix) to get the expectation of the positive patient:

$$E(P) = \frac{kp}{1-q^k} \tag{4.31}$$

In this way, we can also get the expectations of negative patients:

$$E(N) = k - E(P) = k(1-p)\frac{1-q^{k-1}}{1-q^k} \tag{4.32}$$

Analyze the four situations considered above separately, and only when the probability p_{21}, p_{22} occurs, a false positive event may occur. Therefore, using the definition of the false positive rate above, we can get:

$$fpr = \frac{p_{21}k\alpha + p_{22}E(N)\alpha}{k(1-p)} = \alpha - \alpha q^{k-1} - \alpha^2 + 2\alpha^2 q^{k-1} \tag{4.33}$$

Next, we want to observe the relationship between false positive rate and detection error rate, and group size. With a fixed positive rate p, for different detection error rates α, draw a function image of fpr and k (also mark $fpr = \alpha$ this horizontal line)

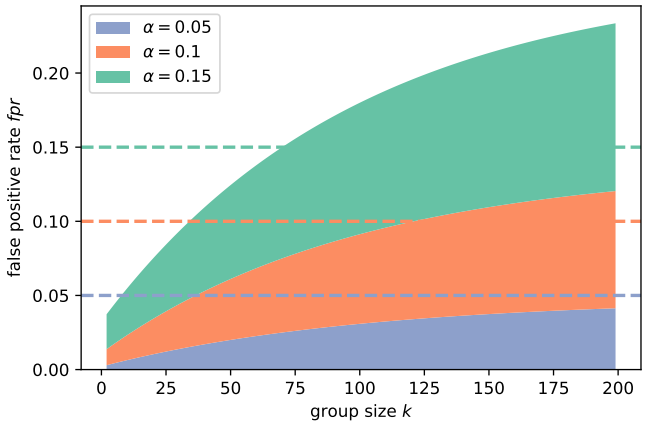


Figure 4.11: the relationship between fpr and group size k and test error α

The dotted line represents the false positive rate without group detection. As can be seen from the above figure, as the packet size increases, as the detection error rate increases, the false positive rate increases, which is also in line with reality.

In the same way, we can calculate the false negative rate:

$$fnr = \frac{p_{12}E(P) + p_{22}E(P)\alpha}{kp} = \alpha(2-\alpha) \tag{4.34}$$

We were surprised to find that the false negative rate has nothing to do with the group size k!

4.2.3 Comprehensive model

Now we want to make m, fpr, fnr as small as possible, and the independent variable is k as integers.

We know that it is often impossible to find a group size k that makes the three smallest at the same time. For multi-objective planning, considering practical problems, it is often to control

the false negative rate and false positive rate to be less than a certain threshold first, and then make the total number of tests as small as possible, that is:

$$\text{s.t.} \begin{cases} \min\limits_{k} m \\ fpr \leq \kappa_1 \\ fnr \leq \kappa_2 \end{cases} \tag{4.35}$$

Note that the false negative rate has nothing to do with the group size, and only depends on the detection error rate α, so there is actually only one constraint above.

For constrained optimization problems, our general approach is to transform them into unconstrained problems: penalty function method.

Take the sufficiently large M_1 and M_2,

$$\begin{aligned} &\min_k m + M_1 \max(fpr - \kappa_1, 0) + M_2 \max(fnr - \kappa_2, 0) \\ \Longleftrightarrow &\min_k m + M_1 \max(fpr - \kappa_1, 0) \end{aligned} \tag{4.36}$$

Below we only need to treat it as an unconstrained optimization problem and draw the relationship between the objective function and the group size.

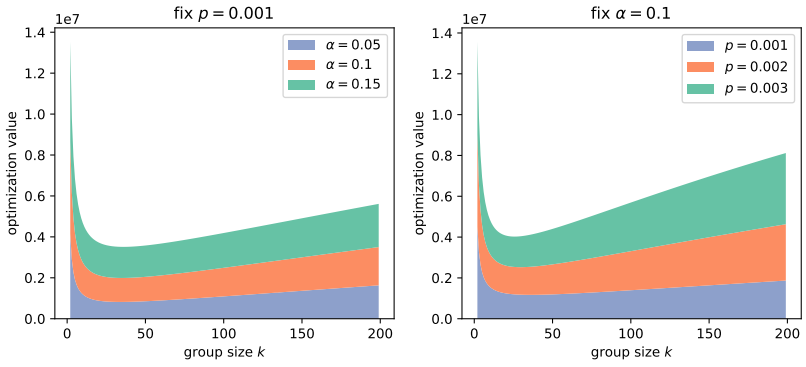


Figure 4.12: the relationship between objective function and group size k

The picture on the left corresponds to the case of different detection error rates α, and the picture on the right corresponds to the case of positive rates p in different populations. It can be seen that as the positive rate p in the population increases, the optimal group size is also larger, which shows the rationality of the results.

4.3 Model with asymptomatic infected people

4.3.1 Social background of asymptomatic infection

Now, COVID-19 is not only appearing in people with cough and fever, but also some asymptomatic people could also be infected. They are a great threat to society as a whole because they can easily infect large groups of people. These people don't have any cough, cold, or fever, which means they're not clinically distinguishable from the rest of the population. The only way to identify them is through a nucleic acid test, which means the rate of false negative rate will depend a lot on error rate α, therefore it is crucial to minimize α. However it is difficult to decrease α and it couldn't be solved in a short time.

As testing for the new coronavirus becomes more common in the United States, scientists have issued a warning about false negatives: Many people who test negative may actually be infected with the virus. An article published in The New York Times by Harlan Krumholz, a professor of medicine at Yale University, pointed out that the error rate of coronavirus testing may not be low, and if the person tested is positive, it is almost certain that the person is infected. Virus, but if it is negative, "maybe you were not infected when the sample was collected." The key word here is "probably". Krumholz said that the testee has been infected, but false negatives that show no infection seem to be a very common phenomenon. The expert went on to point out that there are very few published data on these false negative rates in clinical practice. Research from China shows that the false negative rate may be around 30%[3]. Some American laboratory medicine experts worry that the false negative rate in the United States may be higher.

4.3.2 Problems faced by group test

Although group testing can significantly reduce the number of tests and improve the test efficiency, the false negative rate will also increase as can be seen in the following graph:

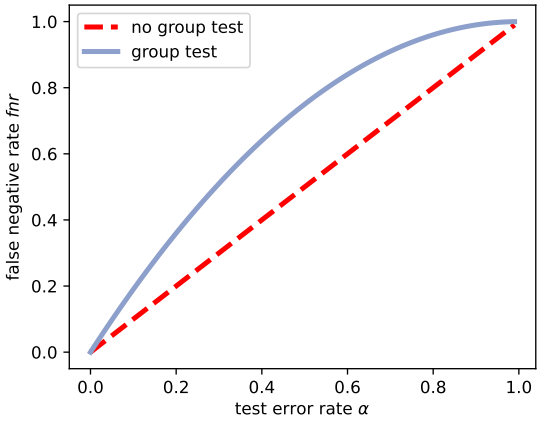


Figure 4.13: group test vs. no group test

The false-negative rate $fnr = \alpha$ when the group is not tested, and the false-negative rate has nothing to do with the group size k when the group is tested, $fnr = 2\alpha - \alpha^2$, use the basic Knowledge of the quadratic function, we know that there is always

$$2\alpha - \alpha^2 > \alpha \qquad \forall \alpha \in [0, 1] \tag{4.37}$$

4.3.3 Solutions for asymptomatic infections

Intuitively, in order to overcome the consequence that group detection will increase the false negative rate, we should increase the number of tests to reduce the false negative rate.Therefore, we decided to repeat the test in the first phase of the test, and then test the positive group one by one.

Assuming that k samples need to be tested, each mixed test needs to be repeated N times, and the other conditions are the same as above.

If the mixed samples appears to be negative which means no more detection is needed, then two kinds of condition are provided.

In the first case, there is no error in the test, there is no positive in the mixed sample and people get negative reports, and in that case the probability is p_{11}

$$p_{11} = (1-p)^k(1-\alpha)^N \tag{4.38}$$

In the second case, a false negative occurs because a negative report is issued while a positive sample is in the mix, and the probability of this is p_{12}

$$p_{12} = (1-q^k)\alpha^N \tag{4.39}$$

If the mixed sample shows positive, then each person in the group needs to be tested, two conditions are also provided.

The first is when the mix is negative, but people get a positive report, which is a false positive, with a probability of p_{21}

$$p_{21} = \sum_{i=1}^{N}(1-p)^k\alpha^i(1-\alpha)^{N-i} \tag{4.40}$$

The second is the error-free test, where there are actually positive cases in the mix, and people actually get positive reports, and the probability is p_{22}

$$p_{22} = \sum_{i=1}^{N}(1-q^k)(1-\alpha)^i\alpha^{N-i} \tag{4.41}$$

Below is method we use to calculate the false negative rate:

$$fnr = \frac{p_{12}E(P) + p_{22}E(P)\alpha}{kp} = \alpha(1-\alpha^N+\alpha^{N-1}) \tag{4.42}$$

The figure below corresponds to the result when the number of repeated tests $N = 1, 2, 3$ for a mixed test sample. It can be seen that the greater the number of repeated tests N, the lower the false negative rate:

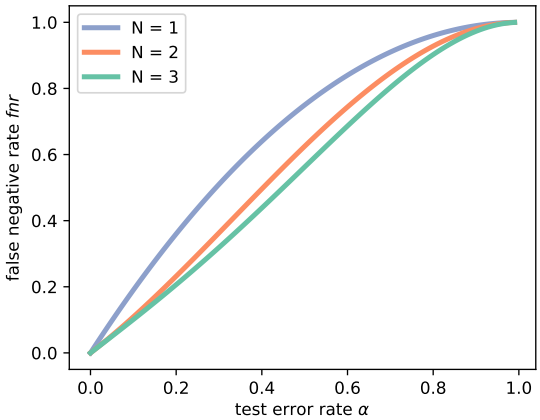


Figure 4.14: The relationship between the number of tests N in mixed tests and the false negative rate

4.4 New detection scheme

4.4.1 Overlap detection

A new grouping scheme will be proposed in this chapter to improve the detection efficiency, and overlap detection will be used in this chapter. Let's start with a two-stage model, where all the variables are the same.But we need to assume that k is a square to make sure that we can arrange k samples into a square.

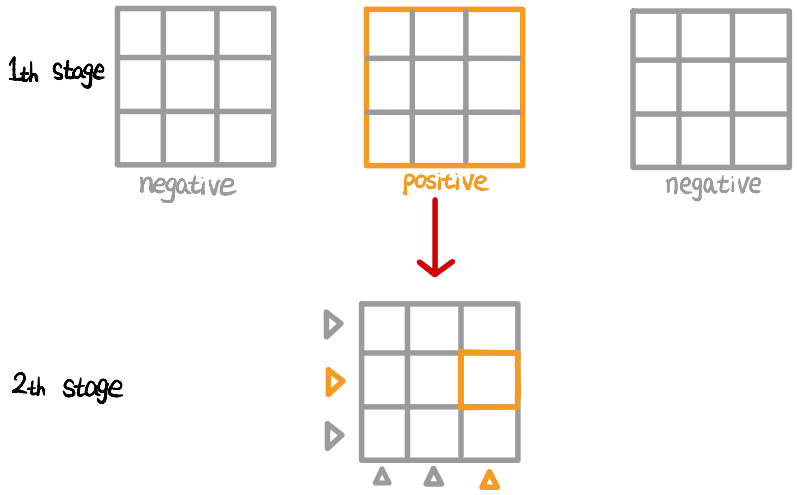


Figure 4.15: Improved group test scheme

As shown in the figure, there are three groups in the first stage, and the second group is positive for nucleic acid test, so in the second stage, we will test the nine people in the second group and find the positive ones. Instead of nine people going through them one by one, we could actually square nine people and test them in rows and columns, and that would only take six tests, but each of the nine people would be tested twice.So since there is no one to check, how to distinguish which person is positive? As shown in the figure, the third column and the second row are positive. If you find the overlap between them, you will find the person who is positive.

In this way, the model can be made more efficient. It used to take 9 times for each person to test twice, but now it can be completed only 6 times. However, this method can only work well when the positive rate p is small. If the positive rate is relatively large, it may overlap with many people, thus increasing the occurrence probability of false positives. However, it is not a big problem, because the impact of false positives on the epidemic is not really great.This method can not only be used for two-dimensional square, but also for three-dimensional cube and even four and higher dimensions.

For higher dimensions, the mathematical description is shown below:

When the mixed sample shows negative with no need to test again, the probability is p_1, the same as (4.1). If the mixed sample shows positive, divide it into d dimensional hypercube, then we need to do $dk^{\frac{1}{d}}$ tests, and the probability is p_2, the same as (4.2). Therefore the average number of tests required for a mixed sample, the expectation is m_1

$$m_1 = p_1 \times 1 + p_2 \times (dk^{\frac{1}{d}} + 1) = q_k + (1 - q^k)(dk^{\frac{1}{d}} + 1) = dk^{\frac{1}{d}} + 1 - dk^{\frac{1}{d}} q^k \tag{4.43}$$

So from above we could know that the total number of tests we need is m

$$m = m_1 \times \frac{n}{k} = \frac{n}{k}(dk^{\frac{1}{d}} + 1 - dk^{\frac{1}{d}}q^k) \tag{4.44}$$

Now we are going to find the minimum number for m, which the variable is k^d

4.4.2 Numerical Example

According to the reference [4], the total urban population of Hong Kong is $n = 750$ million, and the positive rate is $p = 0.01$.

- If a non-repetitive grouping detection scheme is used, the best grouping size is $k = 11$, and the total number of detections is $m = 1466782$
- If an improved grouping detection scheme is adopted, taking $d = 2$, then the best grouping size is $k = 25$, and the total number of detections is $m = 964922$

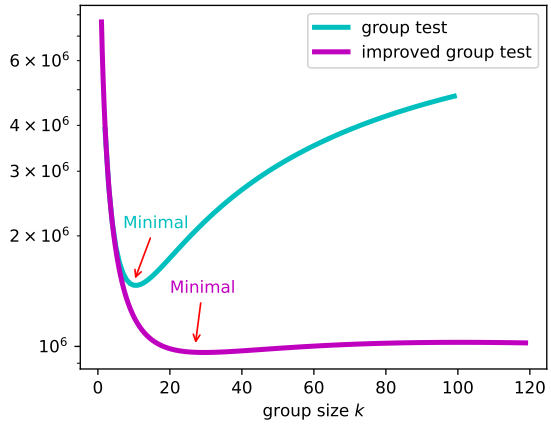


Figure 4.16: Numerical example: shows that the improved packet detection scheme can greatly improve the detection efficiency

5 Sensitivity Analysis

Sensitivity analysis is a method of studying and analyzing the sensitivity of the state or output changes of a system (or model) to changes in system parameters or surrounding conditions. Sensitivity analysis is often used in optimization methods to study the stability of the optimal solution when the original data is inaccurate or is altered.

For this question, both the positive rate p in the population and the error rate α in nucleic acid testing can only be obtained by estimation methods, which are different from the actual values.

So how much influence does the small change of the parameter p and the parameter α have on the group size k?

Note that R_{pk} is the sensitivity of k to the parameter p, and $R_{\alpha k}$ is the sensitivity of k to the parameter α

$$R_{pk} = \frac{\frac{\Delta k}{k}}{\frac{\Delta p}{p}}; \quad R_{\alpha k} = \frac{\frac{\Delta k}{k}}{\frac{\Delta \alpha}{\alpha}} \tag{5.1}$$

The following is the sensitivity of the group size k to the positive rate p and the test error rate α in the population.

p	0.0005	0.00075	0.001	0.00125	0.0015
k	51	41	36	32	30
R_{pk}	0.83	0.56	\	0.44	0.33

α	0.0005	0.00075	0.001	0.00125	0.0015
k	34	35	36	37	39
$R_{\alpha k}$	0.11	0.11	\	0.11	0.16

It can be seen that the sensitivity of p to the parameter, k to the parameter p is not high, and the sensitivity of k to the parameter α is not high, indicating that the model is very stable. In comparison, the group size is more sensitive to the parameter p.

6 Strengths and Weaknesses

6.1 Strengths

- **Considerable**
  Dear Department of Health of the Government of Hong Kong,

  We have comprehensively considered the two-stage and multi-stage models, with error rates and no error rates, and analyzed various situations, which is strong in application.

- **Visualization**
  There are many diagrams to illustrate the relationship between variables, so that you can clearly see the dependence of one variable on another.

6.2 Weaknesses

Since conditional probability analysis is more difficult in multiple stages, we only did upper bound analysis, maybe there are other better ways to deal with it. It may be better to treat the error rates of false positives and false negatives separately

References

[1] Fauci A S, Lane H C, Redfield R R. Covid-19—navigating the uncharted[J]. 2020. Mathematical Society and Addison-Wesley Publishing Company , 1984-1986.

[2] Andreadakis Z, Kumar A, Román R G, et al. The COVID-19 vaccine development landscape[J]. Nature reviews. Drug discovery, 2020, 19(5): 305-306.

[3] `https://www.dw.com/zh/`

[4] `https://www.reuters.com/article/`

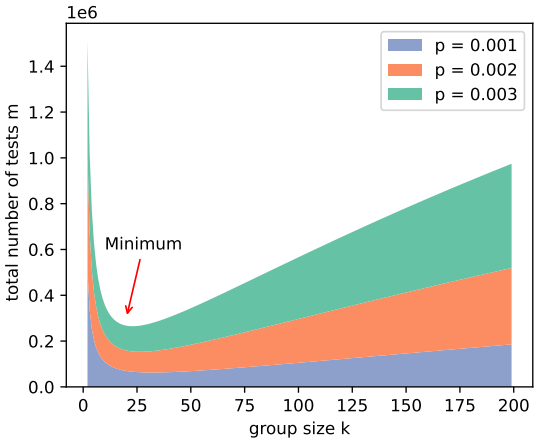
[5] Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 pathophysiology: A review[J]. Clinical immunology, 2020, 215: 108427.

A letter to Hong Kong government's department of health

Dear Sir/Madam:

We are very honored to be able to introduce to you the results of our team's group testing for COVID-19. Minimizing testing number for COIVD-19 can be fairly important in controlling it. As you might have seen, in a situation when only limited testing power is usable, especially where the disease is spreading widely, inability of an overall testing could be very dangerous. As group testing, for your reference, is one of the possible answers, we modeled to find out the most economical solution in terms of it.

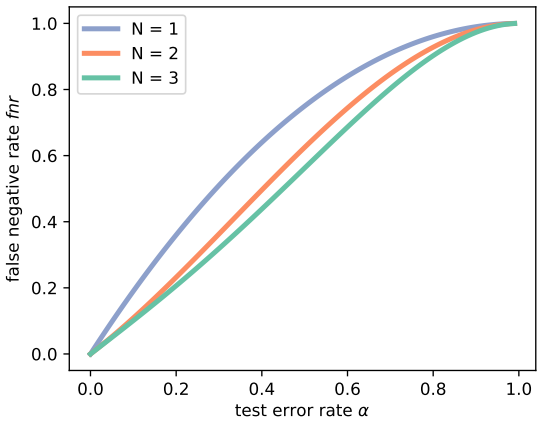
The first part of the model depict the most cost effective solution assuming no error tests is taken place. Under this premise, the model is designed for various stages, which is two, three and various stages, and you will see all those those cases depicted in respective order in the following. The least number of testing necessary in two stages, regardless of the positive rate among the population, is at around a group number of twenty, though the number of tests actually increases in the same level of group size. In three stages, optimal numbers of testing will be reached when the first group's consists of 340 people and 97 people in the second group: those group reported a positive result in the first round of testing. For multiple stages, though you might think that as a nice idea on the first hand, is neither practical nor capable of decreasing testing number according to our verification.



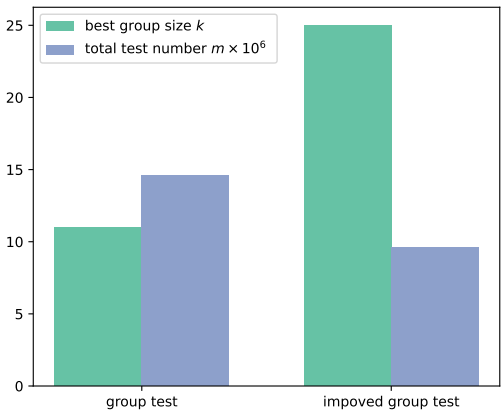
As the first part actually based on the reservation of no mistake made in the testing, which is, unfortunately, impossible to happen, we took testing error into account in the second part of our model– what is way more applicable than the previous one for your reference. In terms of numbers of tests, with the increase of positive rate, test error will also increase, which is something that you might rated as commonsense. Conversely, what contradict the commonsense it the false negative rate does nothing to the group size. The final conclusion reached based on the above statement is when the positive rate increases, increasing the group size can will be a better choice.

To put the problem into a broader context, as you might experience, a person with a clear symptom of covid are more likely to be suspected and in turn isolated. However, as a cunning virus, covid might not present a visible symptom at the first hand. This raises challenge to the group testing method as the error rate will increase comparing to directly testing everyone.

For your understanding, imagine one entire group that is filled with positive patients that was reported negative. In order to solve this, we implemented the method of repeating tests (test N times) at the first hand, which not only decrease the error but also not affecting testing number as well. This turned out to be an effective method.



Possibility of further decrease the testing number actually exists if testing samples overlap with each other. Take Hong Kong, a city you possibly cares the most, as an example. With a population of seven point five million and, supposed, positive rate of point one, our improved method can decrease point five million times of testing, which is significant.



In short, group detection can effectively improve the efficiency of nucleic acid detection and is the key to preventing and controlling the spread of the epidemic. I hope that the Hong Kong government can put forward a suitable plan in light of the actual situation based on the consideration of our research results. We firmly believe that under the leadership of the Hong Kong government, Hong Kong will definitely overcome the epidemic!

Yours sincerely,

Team #1234

Appendices

1 Conditional Expectation Calculation

Problem Suppose the positive rate in the population is p, and the group size is k. It is known that there are positive patients in these k individuals. Let the random variable P represent the number of positive patients in the mixed test sample, and find the expectation of the random variable P.

Solution: Since it is known in advance that there are positive patients in this group of k individuals, at this time, it cannot be directly calculated by the expectation formula of the binomial distribution kp. Need to consider conditional probability.

B: Among the k individuals, i are positive patients

A: There are positive patients among k individuals

Need to calculate the conditional probability $P(B|A)$

$$P(P=i)=P(B|A)=\frac{P(AB)}{P(A)}=\frac{C_k^i p^i(1-p)^{k-i}}{\sum_{i=1}^{k} C_k^i p^i(1-p)^{k-i}}=\frac{C_k^i p^i(1-p)^{k-i}}{1-(1-p)^k}=\frac{C_k^i p^i(1-p)^{k-i}}{1-q^k}$$

Compute the expect:

$$E(P)=\sum_{i=1}^{k} iP(P=i)=\frac{kp}{1-q^k}$$

In the same way, the expectation of the number of negative patients N is

$$E(N)=k-E(P)=k(1-p)\frac{1-q^{k-1}}{1-q^k}$$

2 Two-stage model

```
import numpy as np
n = 1e6
p = 0.001
m1 = []
for k in np.arange(2,200):
    m1.append(n*(1+1/k-(1-p)**k))
m1 = np.array(m1)
p2 = 0.002
m2 = []
for k in np.arange(2,200):
    m2.append(n*(1+1/k-(1-p2)**k))
m2 = np.array(m2)
p3 = 0.003
m3 = []
for k in np.arange(2,200):
    m3.append(n*(1+1/k-(1-p3)**k))
m3 = np.array(m3)
import matplotlib.pyplot as plt
set_figsize()
k = np.arange(2,200)
labels = [f'p = {p}',f'p = {p2}',f'p = {p3}']
```

```
colors = ["#8da0cb","#fc8d62","#66c2a5"]

plt.stackplot(k,m1,m2,m3,labels = labels, colors = colors)
plt.annotate('Minimum',
             xy = (20,0.3*1e6),
             xytext=(20-10,0.6*1e6),
             # weight = 'bold',
             #color= 'b',
             arrowprops = dict(arrowstyle='->',connectionstyle='arc3',color = 'r')
             )
plt.xlabel('group size k')
plt.ylabel(' total number of tests m')
plt.legend()
plt.savefig('m vs k for 2 stages.pdf')
plt.show()
```

3 Approximate relationship

```
p_list = [0.1,0.01,0.001]
y1_list = np.zeros((3,48))
y2_list = np.zeros((3,48))
for i in range(3):
    for k in range(2, 50):
        y1_list[i,k-2] = (1-p_list[i])**k
        y2_list[i,k-2] = 1-k*p_list[i]
k = range(2, 50)
set_figsize((15, 4))
plt.subplot(1,3,1)
plt.plot(k, y1_list[0],c='c', lw = 3, label=r'$(1-p)^k$')
plt.plot(k, y2_list[0], c='m',lw= 3, label = r'$(1-kp)$')
plt.title('$p = 0.1$', fontsize = 18)
plt.xlabel('group size k', fontsize = 18)
plt.ylabel('y', fontsize = 18)
plt.legend()

plt.subplot(1,3,2)
plt.plot(k, y1_list[1],c='c', lw = 3, label=r'$(1-p)^k$')
plt.plot(k, y2_list[1], c='m',lw= 3, label = r'$(1-kp)$')
plt.title('$p = 0.01$', fontsize = 18)
plt.xlabel('group size k', fontsize = 18)
plt.ylabel('y', fontsize = 18)
plt.legend()

plt.subplot(1,3,3)
plt.plot(k, y1_list[2],c='c', lw = 3, label=r'$(1-p)^k$')
plt.plot(k, y2_list[2], c='m',lw= 3, label = r'$(1-kp)$')
plt.title('$p = 0.001$', fontsize = 18)
plt.xlabel('group size k', fontsize = 18)
plt.ylabel('y', fontsize = 18)
plt.legend(fontsize = 18)

plt.savefig('appro relation.pdf')
plt.show()
```

4 Three stage model

```
n = 1e6
def m1_func(p, k1, k2):
    q = 1-p
    m1 = 2*q**k2*(1-q**(k1-k2))/(1-q**k1)+(2+k2)*(1-q**k2*(1-q**(k1-k2))/(1-q**(k1)))
    return m1
def m1_func2(p,k1,k2):
    q=1-p
    p_prime = p/(1-q**k1)
    m1 = (1-p_prime)**k2*2+(2+k2)*(1-(1-p_prime)**k2)
    return m1
def m2_func(m1, p, k1, k2):
    m2 = (m1*k1/k2+(1-p)**k1)*n/k1
    return m2
p = 0.001
m2_1_list = []

for k1 in range(20,1000,10):
    k2 = k1/10
    m1 = m1_func(p, k1, k2)
    m2_1_list.append(m2_func(m1, p, k1, k2))
m2_2_list = []
p = 0.002
for k1 in range(20,1000,10):
    k2 = k1/10
    m1 = m1_func(p, k1, k2)
    m2_2_list.append(m2_func(m1, p, k1, k2))

m2_3_list = []
p = 0.003
for k1 in range(20,1000,10):
    k2 = k1/10
    m1 = m1_func(p, k1, k2)
    m2_3_list.append(m2_func(m1, p, k1, k2))
% set_figsize((7,6))
plt.semilogy(range(20,1000,10),m2_1_list, c='#8da0cb', label = r'p=0.001',lw = 3)
plt.semilogy(range(20,1000,10),m2_2_list, c='#fc8d62', label = r'p=0.002', lw =3)
plt.semilogy(range(20,1000,10),m2_3_list, c='#66c2a5', label = r'p=0.003', lw =3)
plt.xlabel(r'group size $k_1$')
plt.ylabel(r' total number of tests $m_2$')
plt.title('method 1')
plt.legend()
# plt.savefig('m vs k for 3 stages.pdf')
plt.show()
```

5 Asymptomatic infection model

```
fnr_list1=[]
fnr_list2=[]
fnr_list3=[]
fnr_list=[]
for alpha in np.arange(0,1,0.01):
    fnr_list1.append(alpha*(2-alpha))
    fnr_list2.append(alpha*(1+alpha-alpha**2))
    fnr_list3.append(alpha*(1-alpha**3+alpha**2))
    fnr_list.append(alpha)
plt.plot(np.arange(0,1,0.01),fnr_list,lw=3,ls='--',c='r',label='no group test')
plt.plot(np.arange(0,1,0.01),fnr_list1,lw=3, c=colors[0],label='group test')
```

```
plt.xlabel(r'test error rate $\alpha$')
plt.ylabel(r'false negative rate $fnr$')
plt.legend()
plt.show()
# labels=[r,r'test number = 2',r'test number = 3',r'test number = 4']
colors = ["#8da0cb","#fc8d62","#66c2a5"]
set_figsize()
# plt.plot(np.arange(0,1,0.01),fnr_list,lw=3,ls='--',c='r',label='no group test = 1')
plt.plot(np.arange(0,1,0.01),fnr_list1,lw=3, c=colors[0],label=r'N = 1')
plt.plot(np.arange(0,1,0.01),fnr_list2,lw=3,c=colors[1],label=r'N = 2')
plt.plot(np.arange(0,1,0.01),fnr_list3,lw=3,c=colors[2],label=r'N = 3')
plt.xlabel(r'test error rate $\alpha$')
plt.ylabel(r'false negative rate $fnr$')
plt.legend()
plt.show()
```

6 Numerical example

```
import numpy as np
n = 7.5e6
p = 0.01
m1 = []
for k in np.arange(2,100):
    m1.append(n*(1+1/k-(1-p)**k))
m1 = np.array(m1)
plt.semilogy(np.arange(2,100),m1,c='c',lw=3)
plt.xlabel(r'group size $k$')
plt.ylabel(r'total number of test $m$')
z = np.where(m1==min(m1))
k_best=z[0][0]
print(f'the best group size is {k_best+2}, and the minimum number of tests is {m1[k_best

n = 7.5e6
p = 0.01
q=1-p
d = 2
m2 = []
kk_list = np.arange(1,11,0.1)
kkk_list = kk_list**d
for k in kkk_list:
    m2.append(n/k*(d*k**(1/d)+1-d*k**(1/d)*q**k))
m2 = np.array(m2)
plt.semilogy(kkk_list,m2,c='m',lw=3)
plt.xlabel(r'group size $k$')
plt.ylabel(r'total number of test $m$')
z2 = np.where(m2==min(m2))
k_best=z2[0][0]
print(f'the best group size is {k_best}, and the minimum number of tests is {m2[k_best]}
plt.semilogy(np.arange(2,100),m1,c='c',lw=3, label='group test')
plt.semilogy(kkk_list,m2,c='m',lw=3,label='improved group test')
# plt.axhline(y=n,ls='--',c='r',lw=3,label='no group test')
plt.annotate(r'Minimal',
            xy = (10,1.5e6),
            xytext=(7,2.1e6),
            # weight = 'bold',
            color= 'c',
            arrowprops = dict(arrowstyle='->',connectionstyle='arc3',color = 'r')
```

```
            )
plt.annotate(r'Minimal',
            xy = (27,1e6),
            xytext=(20,1.4e6),
            # weight = 'bold',
            color= 'm',
            arrowprops = dict(arrowstyle='->',connectionstyle='arc3',color = 'r')
            )
plt.xlabel(r'group size $k$')
plt.ylabel(r'total number of test $m$')
plt.legend()
plt.show()
```