Blood Testing Problem

Yibo Shen 2017011427,Zhenyu Hou 2017011448,Huahao Lu 2017011477
Department of CST, Tsinghua University, 100084

Abstract

It is a critical problem in actual life to select effective screening method when testing the population attached by diseases with low incidence. In terms of such actual circumstance that can be classified into blood test, *Rober Dorfman* raised a testing method which can improve the efficiency remarkably. First, the expected times of comparing for this method is proved mathematically and mathematical analysis is conducted on its efficiency. Subsequently, two improved methods put forward by the successors are studied. We will analyze the actual efficiency from two different perspectives, that is, theoretical demonstration and experimental simulation. Finally, the improving strategy is proposed. Similarly, we have conducted plenty of data simulation based on the method and obtained more accurate estimation results.

Keywords: screening, optimization

Authors: Yibo Shen, undergraduate of Grade 2017, Computer Engineering Department of Tsinghua University (student ID 2017011427)

Zhenyu Hou, undergraduate of Grade 2017, Computer Engineering Department of Tsinghua University (student ID 2017011448)

Huahao Lu, undergraduate of Grade 2017, Computer Engineering Department of Tsinghua University (student ID 2017011477）

1. Introduction

There are tens of thousands of recessive virulence gens in human's gene pool and each combination can cause some certain disease. However, except some common diseases, the incidence rates of most of the diseases are very low, most of which maintain at one ten-thousandth order of magnitudes. (citation can be inserted here). In the case when it has been known that some disease has very low incidence, what test strategy should we adopt to improve the test efficiency and lower the times of comparing to utmost extent?

In actual life, the significance of such optimization problem is obvious. For example, during the Second World War, American military required that all the new recruits should have physical examination to guarantee the health condition of the army. However, since the incidence of many diseases are extremely low, the efficiency of checking one by one would be very low, which could cause a great waste of both financial and material resources to America in war. Therefore, it is of great necessity to put forward the optimization strategies mentioned above.

As early as in 1943, statistician *Rober Dorfman* had raised a classical testing policy, which can significantly improve the testing efficiency. This study is also based on his classical strategy. We will conduct further studies after investigating this classical strategy.

2. Method validty

Let X be the blood test times required by each person of the population, and then the distribution sequence of X is:

X	$\frac{1}{k}$	$1+\frac{1}{k}$
p	$(1-p)^k$	$1-(1-p)^k$

So the average times for each person to test blood is

$$E(X)=\frac{1}{k}(1-p)^k+(1+\frac{1}{k})[1-(1-p)^k]=1-(1-p)^k+\frac{1}{k}$$

It can thus be seen that only when k is selected and have

$$\frac{1}{k}+1-(1-p)^k<1$$

or

$$(1-p)^k>\frac{1}{k}$$

In this way, the times of blood test can be reduced. In addition, k can also be selected appropriately to make the times of blood test reach the minimum value. For example, when p=0.1, for different ks, the values of E(x) are as shown in the following table. It can be seen from the table, when k $\geq$34, the average times of blood test is more than 1, that is, it is larger than the workload of testing blood respectively; however, when k $\leq$ 33, the average times of blood test is reduced to varying degrees. Especially when k=4, the average times of blood test is the least and the workload can be reduced by 40%.

k	2	3	4	5	8	10	30	33	34
E(X)	0.690	0.604	0.594	0.610	0.695	0.751	0.991	0.994	1.0016

We can also calculate the optimal grouping number of people k_0 based on different incidence rate p, as shown in the following table, from which it can be seen that the smaller incidence p is, the larger the efficiency of grouping test is. For example, when p=0.01, if 11 persons are tested in one group, the workload of blood test can be reduced by about 80%.

3. Improvement

In this section, the following two blood test schemes are considered to improve the original scheme:

Scheme 1: For the grouping displaying positive, it is not to test the blood samples of all the members, but to test the blood sample of each member one by one. If the first infected blood sample is tested out, the blood samples of the remaining members are tested in mixture. If the blood sample shows negative, it suggests that all the remaining members are normal; if positive, the blood samples of all the remaining members should be checked one by one. When the next infected blood sample is tested out, test the blood samples of the remaining members in mixture. Repeat the procedure until the mixture of the remaining blood samples displays negative.

Scheme 2: *Dorfman's* strategy can be regarded as a two-stage method: Group first and then examine each individual of the infected group. Three-stage method can be considered, that is, divide first the population into large groups and then the large group displaying positive is divided further into smaller groups. Furthermore, the general s-stage method can be taken into consideration.

Under normal circumstances, the incidence rate of diseases generally ranges from one in thousandth to one in hundred thousandth. In addition, in the simulation and theoretical analysis of both schemes, the total amount of the test samples adopts the scope from hundred thousand to four million and the incidence adopts the scope from one in five thousand to one in hundred thousand. Later, a series of approximation and simulation data analyses are conducted, which relates also with the scope.

3.1 Theoretical analysis of scheme 1

In terms of each large group with positive testing results, it is supposed that the number is n and number of the diseased members is k. Assume that the tested patient ranks at position x, the total times of test that is required is $x+1$.

Suppose the patients are distributed averagely within the large group, the probability for any patient appearing at position y and before satisfies:

$$P(y)=\frac{y}{n}$$

for $\forall y<n$

$$P(y)=(\frac{y}{n})^k$$

there is

$$P(x=y)=p(\frac{y}{n})^{(k-1)}$$

So, the mathematical expectation of test times t satisfies:

$$E(t)=\int_0^N Np(\frac{y}{n})^k dy+k=\frac{k}{k+1}n+k$$

in case the number n of the large group satisfies

$$n=\frac{m}{\sqrt{p}}$$

the expectation probability of having disease of each large group is np, we have
$E(np)=m\sqrt{p}<<1$

[1]To further solve the expectation of group with k members, we have k ~ $B(p,n)$,meanwhile it satisfies $(1-p)\approx 1$ within the scope of the study.

For $\forall m\in Z^*$, we have

$$\frac{P(k=m+1)}{P(k=m)}=\frac{n-m+1}{m+1}p<np<<1$$

So for large groups whose members can be compared with $\sqrt{\frac{1}{p}}$, if the test result is positive, it can be approximately considered that there is only one diseased people in each group.

So, it can be approximately estimated that the times for each large group with positive test result to be tested is \frac{N}{2}. In this case, the capacity of the large group is also far more larger than 1, so we have.

$$E(t)=\frac{N}{m}\sqrt{p}+\frac{N}{2}m\sqrt{p}$$

The minimum value is taken when m = $\sqrt{2}$. In this way, $t = \sqrt{2}N\sqrt{p}$.

3.2 Simulation of scheme 1

First, I simulated the relation between the scale of the large group and the expectation value of the test under different incidences. m in the scale of large group n=$\frac{m}{\sqrt{p}}$takes $\frac{1}{\sqrt{2}}, 1, \sqrt{2}$. The results are shown in the following figure, in which the horizontal axis is the reciprocal of incidence, the vertical axis is the test times, and the sample size is n=4,000,000. Each incidence is simulated 100 times to take the mean value.

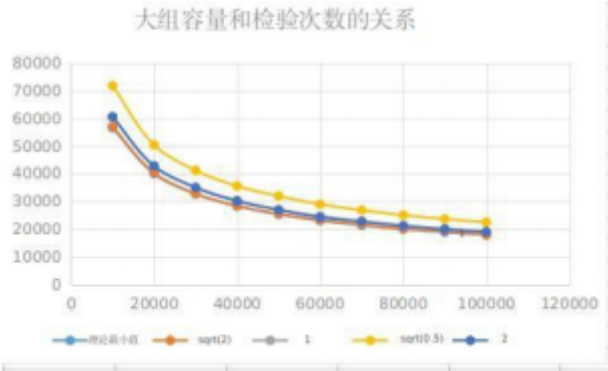

Fig. 1 Relations between the scale of large group and the expectation value of the test under different incidences

It is found within the obtained scope of incidence, the times of test is relatively the least when $m = \sqrt{2}$. In addition, under the capacity of the sample, the simulation results and the theoretical results are approximately identical (indicating the curve of the test times of $m = \sqrt{2}$ completely cover the theoretical curve), which has verified the above theory.

After that, with the constant $m = \sqrt{2}$, we simulated the relations between the sample capacity and the expectation value of the test. The horizontal and vertical axes have the same significance as above. The incidence is constant, which is 0.0002. The capacity of each large group is the mean value within 100 simulations.

It can be seen the test times and the scale of the large group linearly conforms well.

3.3 Theoretical analysis of scheme 2

For case 2, we assume first l groups are divided during the first grouping, and l is far larger than Np. Then, the final expectation of the test times is:

$$E(t) = l + Npf(\frac{N}{l})$$

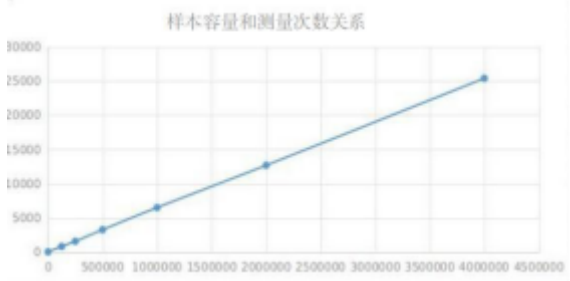

Fig. 2 Relations between the scale of large group and the expectation value of the test under the same incidence

When l is very large.

According to the theory in scheme 1[1], it can be considered that there is at most one patient in each group, when the expectation of the test times is the scale of the second group s where.

$$s = s + \frac{l}{s}$$

If the minimum value is taken, s=$\sqrt{l}$.

The final expectation of the test times is $t = l + 2Npf(\frac{N}{l})$

When t takes the minimum value, it satisfies $\frac{dt}{dl} = 0$, so $l = Np^{\frac{2}{3}}$, and the member number of each group is $p^{-\frac{2}{3}}$. The expectation of the total test times t is $3Np^{\frac{2}{3}}$, which is $1.5p^{\frac{1}{6}}$ times of the original scheme.

When $p = 0.0001$, $1.5p^{\frac{1}{6}} = 0.323$,which is superior to the first scheme discussed above.

In addition, when taking s-stage scheme into consideration further, recursion formula

$$f_n(k) = min(j + f_{n-1}(\frac{j}{k}))$$
$$f_1(k) = k$$

has a solution

$$f_n(k) = n\sqrt[n]{k}$$

In s-stage scheme, in case the number of the groups divided in the first time is j, the test times t satisfies:

$$f_s(t) = j + f_{s-1}(t)(\frac{N}{j})$$

It has the same form as the above recursion formula.

In this way, the optimal grouping number is

$$j = Np^{\frac{s}{s+1}}$$

When the capacity of each group is $n = p^{-\frac{1}{s+1}}$, and the final test times is $t = sNp^{\frac{s}{s+1}}$.

It is found that for each s, there is a real number g, making $s = -\log_g p, t = -gp\log_g p$.

In terms of the partial derivative of t about g, it is found that when $g = e$, t takes the minimum value $-ep\ln p$, so when the cycle index s meets $s = -\ln p$, the expectation of the test time takes the least.

3.4 Simulation of scheme 2

During programming simulation, I found by running the meter that when $p = 0.000001$, $\sqrt[3]{p}$ is already not "far larger" than l, so the approximation result would fail when s is larger or p is larger. For this reason, I only conducted simulation when $s = 2$.

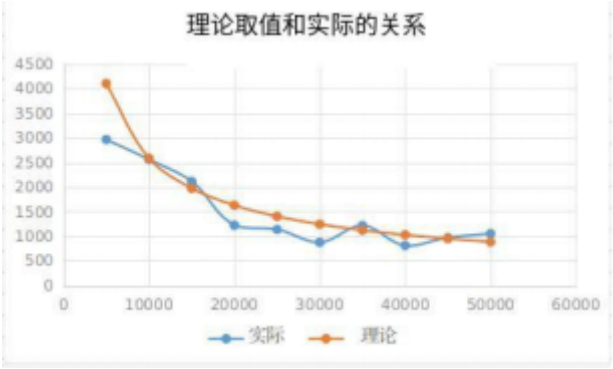

Fig. 3 Relations between the theory and the experimental simulation

It can be found there is obvious deviation between the actual curve and the theoretical curve. Through analysis, it is thought that the reason could be the failure of hypothesis $\sqrt[3]{p} >> 1$. If not, the deviation resulted from approximation failure will become larger. From figure we can get that the actual curve deviates the largest from the theoretical curve at $p = 5000$, which can be a proof of the hypothesis.

After that, I drew a figure to show the relations between the sample capacity and the test expectation under the same incidence. The horizontal and vertical axes have the same meaning as mentioned above, and the incidence rate is constant 0.00002. The data of each group is the mean value of 200 simulation.

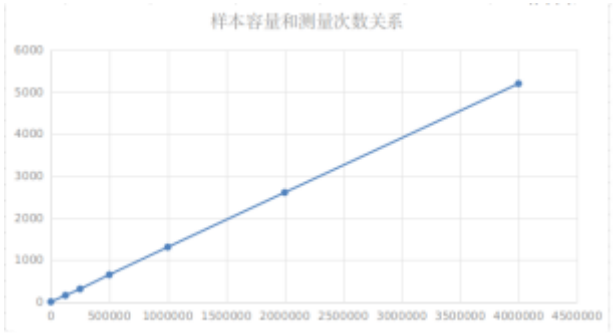

Fig. 4 Relations between the sample capacity and the expectation value of the test

It can be seen that within the scope of the sample capacity that is considered, excellent linearity is retained.

4. New improvement strategy

Inspired by the idea of binary search I have learned in the class of data structure, I designed a method requiring fewer test times than the above methods in practical data range. Specific operation is as follows:

1. Divide equally all the samples into k groups, each of which is tested in mixture.
2. Take those with positive test results. If there is more than one positive sample, implement procedure 1 (i.e., divide them in to k subgroups for test in mixture) till such condition cannot be satisfied.

Through theoretical analysis of the test times of this scheme, it can be seen easily that if the sample capacity is N, the times of repeated implementation is $\log_k N$. Suppose the incidence of the sample is Np, the number of the groups that require to be tested in each cycle is no larger than kNp.

Considering that only two groups are in need be tested for the first time, the total test times that is required is strictly smaller than $kN \log_k N$.

When $k=2, n=4000000, p=0.0002$, here comes:

$$\begin{aligned} t &= 2Np\log_2 N \\ &= 3.5\times 10^4 \\ &< 4.1\times 10^4 \\ &= 3Np^{\frac{2}{3}} \end{aligned}$$

So, this new scheme has obvious advantages theoretically.

In addition, when working out k which makes the theoretical upper bound the smallest, function $g(x)=k\log_k 2$ is considered, and it is found the value is the least when $k=3$. However, whether the actual test times is the least when $k=3$ still requires further analysis.

4.1 Data simulation

When simulating, I firstly tested whether the upper bound theory when $k=2$ accords with the reality. The results are shown below. In the figure, the horizontal axis is the incidence, the vertical axis is the times of testing and the sample capacity $N=4000000$. The data of each group is simulated 200 times to take the mean value.

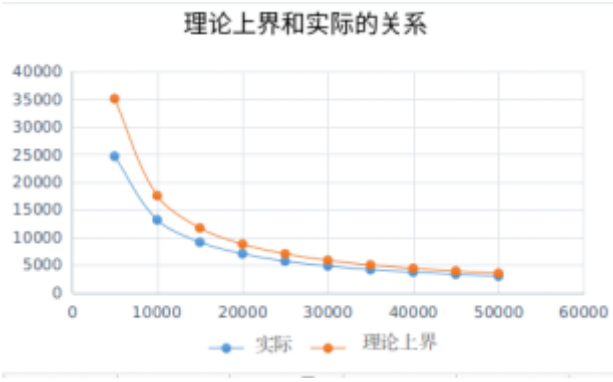

Fig. 5 Relations between the sample capacity and the expectation value of the test

No matter how the incidence changes, the red curve in the figure remain below the orange curve, indicating the upper bound exists strictly.

Based on this, I simulated the efficiencies of the proposed method and the two methods discussed above with large sample capacity ($N=4000000$) and small capacity ($N=100000$).

It is found that under the premise of $0.000002\le p\le 0.0002$, no matter with large or small sample capacity, the efficiency when $k=2$ obviously outperforms the other two.

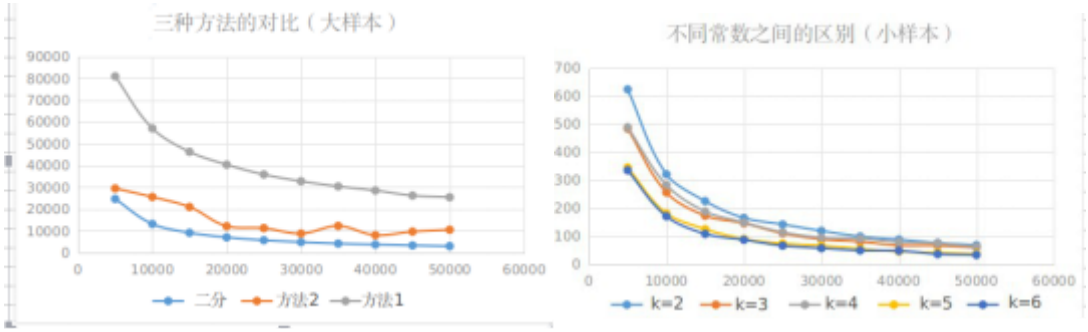

Fig. 6 Comparison of three methods under the same sample capacity

To test the conditions of different grouping, I conducted the simulation with $k=2,3,4,5,6$, as shown in the following figure.

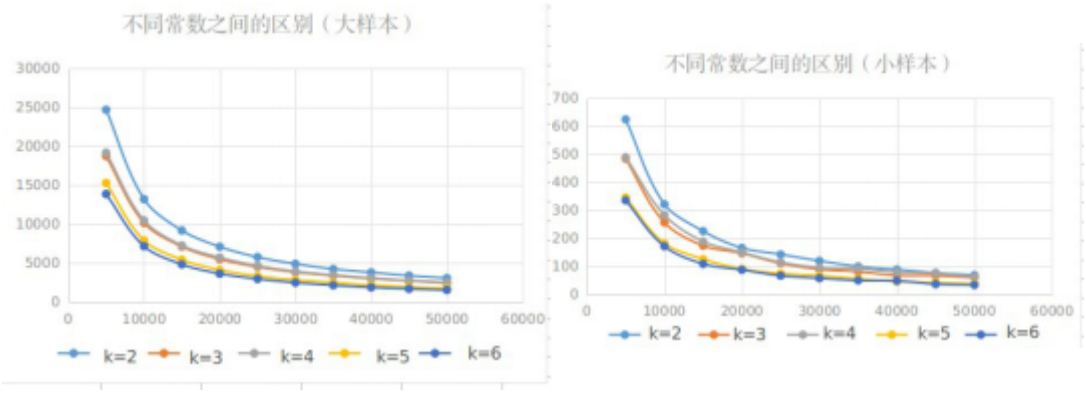

Fig. 7 Comparison of different grouping numbers under different sample capacities

It is found the actual sequence of the test times is not exactly the same as that of the upper bound. Within the simulated incidence scope, the result when $k = 4$ is close to that when $k = 3$, while the result when $k = 6$ is close to that when $k = 5$ but smaller than the former. It is analyzed that the search depth becomes smaller with the increase of k, so that the proportion between the mean value of the positive group and the total number of the patients that are found in each layer decreases. Thus, the upper bound becomes "wider", which cannot reflect the actual expectation value of the test.

5. Conclusion

Blood test problem is an application-based problem with large sample and small probability. When designing a test strategy, it is necessary to approximate the probability distribution and then optimize it from different directions.

Through theoretical analysis and simulated experiment, the following conclusions can be obtained:

(1) Under the basic strategy of grouping test, the optimal strategy is to have the number of people in each big group should be $\frac{1}{\sqrt{p}}$, and the expectation value of the test times should be $E(t) = 2N\sqrt{p}$.

(2) Both of the given methods can optimize the results of obtained by (1) within the actual scope.

For strategy 1, it should be noted the number of people in each big group should be changed into $\sqrt{\frac{2}{p}}$ so that the expectation value of the test times can be optimized to $0.5\sqrt{2}$ times than that of the grouping test.

For strategy 2, when grouping for s times, it should be noted that the average number of person grouped in the k_{th} time should be $p^{-\frac{s-k}{s}}$ so that the expectation value of the test times can be optimized to $(s+1)p^{0.5-\frac{1}{s+1}}$ times than that of the origin.

(3) The idea of k grouping is generated from binary method. In this method, the upper bound of the test times is strictly $kNp\log_k N$, which is smaller than the expectation values of the previous two strategies. However, with the increase of k, the number of recursion layers decreases, so the expectation value of the test times does not increase monotonically with k. Yet in actual test, the larger k is, the farther the distance between the test number and the theoretical upper bound becomes.

References

[1] The Blood Testing Problem, H. M. Finucan, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 13, No. 1 (1964), pp. 43-50