

The Blood Testing Problem

Suppose that you have a large population that you wish to test for a certain characteristic in their blood or urine (for example, testing all NCAA athletes for steroid use or all US military personnel for a particular disease). Each test will be either positive or negative.

Since the number of individuals to be tested is quite large, we can expect that the cost of testing will also be large. How can we reduce the number of tests needed and thereby reduce the costs? If the urine could be pooled by putting a portion of, say, 10 samples together and then testing the pooled sample, the number of tests might be reduced. If the pooled sample is negative, then all the individuals in the pool are negative, and we have checked 10 people with one test. If, however, the pooled sample is positive, we know only that at least one of the individuals in the sample will test positive. Each member of the sample must then be retested individually and a total of 11 tests will be necessary to do the job. The larger the group size, the more we can eliminate with one test, but the more likely the group is to test positive.

Would pooling 5 specimen be better than pooling 10? What is the relationship between the probability of an individual testing positive and the group size that minimizes the total number of tests required? Certainly, we anticipate that the larger the probability of an individual testing positive the smaller the group size, while the smaller the probability, the larger the group size required.

Problem Statement

You have a large population (N) that you wish to test for a certain characteristic in their blood. Each test will be either positive or negative. Since the number of individuals to be tested is quite large, you wish to reduce the number of tests needed to screen everyone and thereby reduce the costs. If the blood could be pooled by putting G samples together and then testing the pooled sample, the number of tests required might be reduced. What is the relationship between the probability of an individual testing positive (p) and the group size (G) that minimizes the total number of tests required? Use your solution to determine the number of tests required to find 100 individuals who will test positive in a population of 1,000,000.

The Basic Model

An essential aspect to developing any model is to consider the simplest case that embodies the essence of the problem. For the group testing problem, this is a solution that uses only one group test, and then tests everyone remaining individually. If the students cannot solve this problem, they will not be able to solve a more involved model that is perhaps more realistic. Further, the solution to the simplest situation often is helpful in arriving at a more general solution. What follows is the approach taken by most student groups.

Since there are N people to be tested in groups of size G , the initial number of tests needed to test these group is $\frac{N}{G}$. The probability of testing positive is p , so we expect that Np people will test positive. With the worst case assumption that exactly one person in each group will test positive, this means that Np of the groups will test positive, and NpG people remain to be tested. The number of tests needed for this testing protocol is given by

$$T(G) = \frac{N}{G} + NpG.$$

In the worst case, if any group tests positive, all members of the group will test positive. So the number of tests needed to test using one group test and then testing everyone remaining individually is

$$T = \frac{N}{G} + NpG = N\left(\frac{1}{G} + pG\right).$$

Since the factor N produces a vertical stretch in the graph of T , the value of G that minimizes the number of tests will not be affected by N , so we set $N=1$ for convenience. For a given value of p , we can determine the group size G which minimizes the number of tests, and therefore the costs, by using a graphing calculator to zoom and trace. For, example, if $p = 0.01$, we can find the minimum value on the graph of $T = \frac{1}{G} + (0.01)G$ as shown in Figure 1 below.

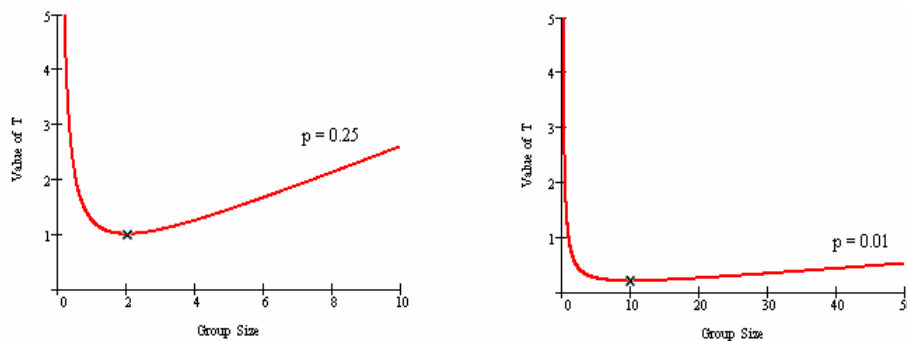


Figure 1: Graphs of $T = \frac{1}{G} + pG$ for $p = 0.25$ and $p = 0.01$

Repeating the process for different values of p , we generate the table below:

p	0.25	0.20	0.15	0.10	0.05	0.03	0.01	0.005	0.001	0.0005	0.00001
G	2.0	2.24	2.58	3.16	4.47	5.77	10.0	14.14	31.62	44.72	100.0

By using techniques of data analysis on the scatterplot for this data, we can create a model relating the best group size G to the probability p .

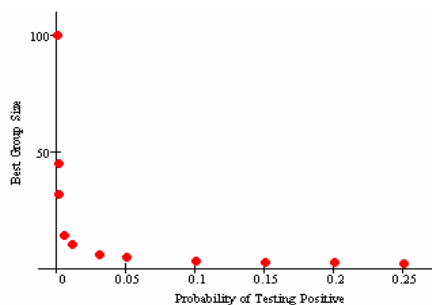


Figure 2: Scatterplot of “Best Group Size” vs Probability

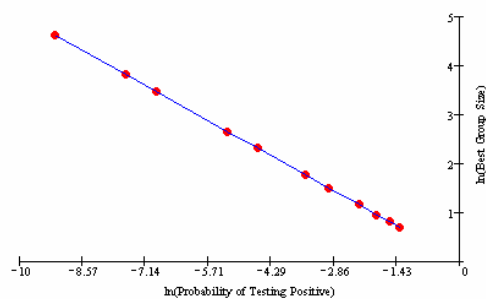


Figure 3: Log-log Re-expression to Linearize Data

By re-expressing the data with a log-log plot, we linearize the data. The least-squares equation is $\ln(G) = 0.00005 - 0.5 \ln(p)$, so $G = \frac{1}{\sqrt{p}}$. If $G = \frac{1}{\sqrt{p}}$, the total number of tests needed is

$$T = N \left(\frac{1}{G} + pG \right) = N \left(\sqrt{p} + \frac{p}{\sqrt{p}} \right) = 2N\sqrt{p}.$$

In our example, 100 individuals testing positive can be found in a population of one million people in approximately 20,000 tests. Further, if $p \geq \frac{1}{4}$, then $2N\sqrt{p} \geq N$, and it is counterproductive to test in groups.

Refining and Improving the Model

Of course, there is no reason to re-test everyone individually. Since G is independent of N , we could retest all of the NpG remaining after the first group test in similar groups. We already know that $G = \frac{1}{\sqrt{p}}$. However, since we have already eliminated a large number of people in the first phase of testing, the value of p will be much larger for the second group test. There are Np people that we expect to test positive and NpG people remaining to be retested. The probability of testing positive in the second round is $p^* = \frac{Np}{NpG} = \frac{1}{G} = \sqrt{p}$. So the next test

should be done with $G = \frac{1}{\sqrt{p^*}} = \frac{1}{\sqrt[4]{p}}$.

Continuing in this fashion, we find the group sizes to be

First Grouping	$\frac{1}{\sqrt{p}}$	New Probability	\sqrt{p}
Second Grouping	$\frac{1}{\sqrt[4]{p}}$	New Probability	$\sqrt[4]{p}$
Third Grouping	$\frac{1}{\sqrt[8]{p}}$	New Probability	$\sqrt[8]{p}$
n th Grouping	$\frac{1}{\sqrt[2^n]{p}}$	New Probability	$\sqrt[2^n]{p}$

Stop grouping when the new probability is greater than 0.25.

When do you stop grouping and test everyone individually? We want to know for what n is $\sqrt[2^n]{p} \geq \frac{1}{4}$. Solving for n , we find that $n = \frac{1}{\ln(2)} \ln \left(\frac{\ln(p)}{-\ln(4)} \right)$.

The iterated model just created works quite well, reducing the number of tests dramatically, even though in practice we violate the assumption upon which the solution was based. In the example of finding 100 positive individuals in a population of 1,000,000, testing in groups of 100, 10, and 3, and then individually requires only 11,634 tests. However, this is clearly not an optimal solution. In creating the model, we assumed that we would group only

once and then retest individually. The group size $G = \frac{1}{\sqrt{p}}$ was determined on the basis of that assumption. Is it possible to determine the number of tests needed by taking the additional group tests into account? A second model extends this initial solution.

The model just created works well, reducing the number of tests dramatically. In the example of finding 100 positive individuals in a population of 1,000,000, testing in groups of 100, 10, and 3, and then individually requires only 11,634 tests.

Round of Testing	1	2	3	4
Number to Test	1,000,000	10,000	1,000	300
Group Size	100	10	3	1
Number of Tests	10,000	1,000	334	300
Number to Retest	100(100)	100(10)	100(3)	0

Calculus Solution:

This is a straight-forward calculus problem for first semester students. We have the same model

$$T = \frac{N}{G} + NpG$$

To determine the best group size G , we differentiate with respect to G .

$$\frac{dT}{dG} = -\frac{N}{G^2} + Np,$$

and find the value of G that makes this derivative zero.

If $\frac{dT}{dG} = 0$, then $G = \frac{1}{\sqrt{p}}$. The rest of the problem proceeds as before, and we find that 11,634 tests are needed to find 100 out of 1,000,000. This is a large reduction, but we can do even better!

The Multiple Group, Iterated Model

As with the initial solution, the starting point is with the simplest model that contains the essence of the problem. In this case, it is a model that allows for two group tests and then testing everyone remaining individually. So,

$$T = \frac{N}{G_1} + \frac{NpG_1}{G_2} + NpG_2.$$

It seems that T is a function of two variables, G_1 and G_2 , which is beyond the scope of an introductory course in calculus. Is possible to rewrite this as a single variable problem? A few hints are generally in order at this point.

With some encouragement, students will realize that they know the solution to the last part of the problem,

$$T = \frac{N}{G_1} + \left[\frac{NpG_1}{G_2} + NpG_2 \right],$$

the problem of minimizing the number of tests with one test and then testing everyone remaining individually. Recall that this group size was independent of the number being tested. So, in fact,

we know $G_2 = \frac{1}{\sqrt{p^*}}$, where p^* is the probability of testing positive after the first test. But p^* is just the expected number testing positive divided by the total number in the present population. So $p^* = \frac{Np}{NpG_1} = \frac{1}{G_1}$. Substituting, we know that $G_2 = \sqrt{G_1}$. The total number of tests can now be written as a function of the single variable G_1 .

$$T(G_1) = \frac{N}{G_1} + 2Np\sqrt{G_1}.$$

Then

$$\frac{dT}{dG_1} = \frac{-N}{G_1^2} + \frac{Np}{G_1^{1/2}}$$

and elementary calculus shows that the optimum value for G_1 is $G_1 = p^{-2/3}$. If two groupings are used, the sizes of the groups should be

$$G_1 = p^{-2/3} \text{ and } G_2 = p^{-1/3}.$$

Extending the grouping to three continues the pattern. If

$$T = \frac{N}{G_1} + \frac{NpG_1}{G_2} + \frac{NpG_2}{G_3} + NpG_3,$$

we know that $G_2 = (p^*)^{-2/3}$ and $G_3 = (p^*)^{-1/3}$, with $p^* = \frac{1}{G_1}$. So $G_2 = G_1^{2/3}$ and $G_3 = G_1^{1/3}$.

Rewriting, we find that

$$T(G_1) = \frac{N}{G_1} + 3NpG_1^{1/3}$$

Again, elementary calculus shows that the optimum group sizes are $G_1 = p^{-3/4}$, $G_2 = p^{-1/2}$, and $G_3 = p^{-1/4}$. Repeating the analysis with four groups generates the optimum group sizes $G_1 = p^{-4/5}$, $G_2 = p^{-3/5}$, $G_3 = p^{-2/5}$, and $G_4 = p^{-1/5}$.

In general, if a total of n groupings are used, the group sizes are given by

$$\begin{aligned} G_1 &= p^{-\frac{n}{n+1}} \\ G_2 &= p^{-\frac{n-1}{n+1}} \\ G_3 &= p^{-\frac{n-2}{n+1}} \\ &\vdots \\ G_n &= p^{-\frac{1}{n+1}}, \end{aligned}$$

with the k th group of size $G_k = p^{-\frac{n-(k-1)}{n+1}}$. No group of students has yet verified this generalized grouping, but they place their faith in the pattern generated from using 1, 2, 3, 4, and 5 groups. Proving the general case seems beyond their present abilities.

Based on the clear pattern developed by initial specific cases, students argue the total number of tests required with n groupings is

$$T = \frac{N}{G_1} + NnpG_1^{1/n}.$$

What number of groupings n is optimum for a given initial probability p ?

If we consider T as a function of n , we find that

$$T(n) = N \left(p^{\frac{n}{n+1}} \right) (1+n).$$

Differentiating, we find that

$$\frac{dT}{dn} = N \left(p^{\frac{n}{n+1}} \right) \left(1 + \frac{\ln(p)}{n+1} \right).$$

The optimal number of groupings is $n = -\ln(p) - 1$. Recall that we are ignoring any non-integer aspects of the problem. The total number of tests required is given by

$$T = Npe(-\ln(p)).$$

If $p = 0.0001$, this is a reduction by a factor of 400. If 100 out of 1,000,000 had the sought for characteristic, they could be found in around 2,500 tests.

Finding the Optimal Group Size

With this value of n , we can also determine the optimum size of the k th group. We know that $G_k = p^{\frac{n-(k-1)}{n+1}}$, with $n = -\ln(p) - 1$, so the k th group should size should be

$$G_k = p^{\frac{-\ln(p)-k}{\ln(p)}}.$$

While most groups leave their expression for the k th group size in this form, $G_k = p^{\frac{-\ln(p)-k}{\ln(p)}}$ can be simplified.

$$\text{If } G_k = p^{\frac{-\ln(p)-k}{\ln(p)}}, \text{ then } \ln(G_k) = \ln \left(p^{\frac{-\ln(p)-k}{\ln(p)}} \right) \text{ and } \ln \left(p^{\frac{-\ln(p)-k}{\ln(p)}} \right) = -\ln(p) - k.$$

So $\ln(G_k) = -\ln(p) - k$ or

$$G_k = \frac{1}{pe^k}.$$

Students are always surprised to see e show up in the solution. Of course, while this theoretical result is pleasing, it may not be realizable, for $G_k = \frac{1}{pe^k}$ may be too many specimen to handle in a single group.

This problem offers many important teaching points both about mathematical modeling and about calculus. The problem of improving the initial solution by violating the assumptions

creates an interesting discussion about mathematical theory and practice, and the importance of a good approximate solution over an ideal unrealizable one. The importance of iterating the model and refining the solution based on prior work is clear and convincing in this setting. No students have found the more sophisticated multiple group solution without first working through the single group solution and being dissatisfied with it. Also, the importance of considering "what question does this new solution ask?" is seen in several places. We obtain a solution, and immediately use it to further the problem. First, we consider T as a function of G , then as a function of n . The conversations surrounding the solution and in the process of solving this problem encourage essential aspects of modeling.

Final Note

The Precalculus modeling technique will also work on the average case scenario. In this situation, the total number of tests is the sum of the initial number of tests $\left(\frac{N}{G}\right)$ and all the retests $\left(\left(1-(1-p)^G\right) \cdot \frac{N}{G} \cdot G\right)$, so the model is $T(G) = \frac{N}{G} + \left(1-(1-p)^G\right) \cdot \frac{N}{G} \cdot G = N\left(\frac{1}{G} + 1-(1-p)^G\right)$. You may want to verify that calculus is absolutely no help here! However, we can create a table for p and best G as before and fit a model to the data. The result is $G = \frac{1}{\sqrt{p}} + 1$.

REFERENCES

1. Dilwyn Edwards and Mike Hamson, *Guide of Mathematical Modeling*, CRC Mathematical Guides, CRC Press, Boca Raton, Florida, 1989, pp.199-208.
2. William Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Ed.* John Wiley and Sons, Inc., New York, 1968, p. 225.
3. Paul L. Meyer, *Introductory Probability and Statistical Applications, 2nd Ed.*, Addison-Wesley Publishing Co., Reading, Massachusetts, 1970, pp. 131-132.