

Problem

Among a group of 1,000,000 individuals, it is known that exactly 100 have syphilis. Unfortunately, the blood tests are very expensive, and it is too expensive to test everyone individually. However, one may take blood samples from any number of people, combine them, and test the combined sample to see if anyone in that group has syphilis. What is the least N you can find, such that you can identify which 100 people have syphilis using at most N blood tests?

Techniques like this were used to find syphilis cases in large populations during WWII[1].

The winner will be whoever submits the solution with the least N .

Solutions

The best student solution was submitted by Ben Burdett, who found the infections using at most 2621 tests. Randy Berta '76 found a solution using 2534 tests. David Stone '91 had a solution which, though it required more tests, is very elegant; with his permission, here is his solution:

I can do it with N no greater than 12,000. Set up a 1000 by 1000 matrix, identifying each individual uniquely with a cell in the matrix. Take blood samples from individuals, split them, and put them in each of 2000 test tubes—that is, the individual at cell $x = 156$, $y = 232$ will have his / her samples in test tube $x156$ and $y232$. Test those 2000 samples. That will produce at most 100 positives on the x -axis and 100 positives on the y -axis, and so at most 10,000 possible intersections. Retest those 10,000 individually.

Total maximum number of tests: 12000. It could be as few as 2000 in the unlikely event that all the positives wind up in the same row or column.

Binary subdivision

One strategy is as follows. To find the infections in a set of two or more people, break the group into two subsets of equal size (or as nearly equal as you can get if there is an odd number of people), and test each subset. Repeat this process for any subset that tests positive: split it in half and test each half. Of course, if a group of size 1 tests positive, you do not need to split it up—you have just found an infected person.

How many will tests will this require? Let us think of the process as proceeding in rounds: in round 1, we test 2 groups of size 500,000; in round 2, we test (up to) 4 groups of size 250,000; etc. In round 20, any remaining groups have size 1 (20 is the least n such that $2^n \geq 1,000,000$), so we will be done at the end of that round. How many tests must be performed in each round? For the first several rounds, the upper bounds on the number of tests are 2, 4, 8,

16, 32, etc. However, in any given round, at most 100 groups will test positive, leaving at most 200 groups to test in the following round. So we never need to test more than 200 groups in a round. In rounds 1–7, the upper bounds are 2, 4, 8, 16, \dots , 128, from round 8 onward, the upper bound will be 200. So, this requires $2 + 4 + \dots + 128 + 13 \cdot 200 = 2854$ tests.

What is wasteful about this approach? The first round of tests is practically useless, because each group of 500,000 is overwhelmingly likely to contain an infected individual. It would make more sense to break into a larger number of smaller groups before testing. Benjamin Burdett started with groups of size 4096 (plus one group of 576 at the end), and also took advantage of the fact that once you have 100 different groups that are positive, you know each of those groups has exactly one infection, and can use fewer tests to finish. (E.g., if you know a group of 128 people has exactly one infection, you only need to test one of the two subsets of 64: if the test is positive, then the infection is in that set of 64, and if it's negative, then the infection is in the other set of 64.)

A lower bound on the number of tests needed

Each test has two possible outcomes: positive or negative. Given any strategy for using at most N tests, there are at most 2^N ways that the tests could come out. In particular, N tests can at best distinguish between 2^N possibilities. For example, 4 tests can distinguish between at most 16 outcomes, so if there are 20 possible outcomes, the number of tests required is at least 5. More generally, if there are p possibilities, then at least $\log_2(p)$ tests will be required (though it might end up requiring even more).

If we count how many possible ways there can be 100 out of 1000000 people infected, then we will be able to get a lower bound on the number of tests needed. Let $\binom{n}{k}$ (“ n choose k ”) denote the number of ways to choose k elements out of a set of n elements. A formula for $\binom{n}{k}$ is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

For example, the number of ways to choose 3 people from a group of 10 is $\binom{10}{3} = \frac{10!}{3!7!} = 120$. In our case, we are interested in $\binom{1000000}{100}$, which is 10662192428510-62012874518533388803111513947244272930391945948085447596313527456513-22388253308751600589673382019458913096053871115851715576076247999258-41433270198940608835213627244565438262233673022772626479799449169047-66429771247815307377515542049884622522011608437597237595775322846223-52566855382432132280834890599732601135978509136592063612245457634409-82111961752912443226568256491091034269861759900523299691641619115198-117741097602556240000.

Calling this number p , we have $\log_2(p) \approx 1468.38$. So, no technique that uses only 1468 tests will work. The best we can hope for is 1469 tests, though this is probably not attainable.

Another solution

If we could cut the number of possible outcomes roughly in half with each test, then we could get close to the theoretical limit of 1469. Here is one such technique.

We start out with 10^6 individuals, we know 100 of them are infected, and—this will seem contrived at first—we know that at least one person out of the first 999901 is infected. We encode this by the triple $(10^6, 100, 999901)$. More generally, before each test, the relevant information we need in order to proceed will be a triple (P, I, k) , indicating that there are P people left to test, I of them are infected, and at least one of the first k of them is infected. We will perform a test that, for some $m < k$, tests the first m people together. If the test comes back positive, then we now know one of the first m people is infected, leaving us with the triple (P, I, m) . If the test is negative, then we forget about those first m people; of the $P - m$ who remain, there are still I infections, and one of the first $k - m$ must be infected, leaving us with the triple $(P - m, I, k - m)$.

If we ever reach a triple $(P, I, 1)$, we know, without having to perform a further test, that the first person is infected. That person is marked as infected and set aside; of the $P - 1$ that remain, there are $I - 1$ infections, and one of the first $P - I + 1$ must be infected (since the first $P - I + 1$ people include all but $I - 2$ individuals). This leaves us with the triple $(P - 1, I - 1, P - I + 1)$. If we ever reach a triple $(P, 0, k)$, then there are no further infections, and we are done. Or, if we ever reach a triple (P, P, k) , we know all remaining people are infected, and need no further tests.

How do we choose the value of m above? Given any triple (P, I, k) , there is a number $C(P, I, k)$ of possible combinations for who is infected; specifically,

$$C(P, I, k) = \binom{P}{I} - \binom{P - k}{I}$$

(the number of ways to have I infections out of P people, minus those cases where all I infections are among the last $P - k$ people). After testing the first m people, we either have $C(P, I, m)$ combinations remaining (when the test is positive), or $C(P - m, I, k - m)$ combinations remaining (when the test is negative). We choose the value of m that makes $C(P, I, m)$ as close as possible to $C(P - m, I, k - m)$ —in other words, we choose the value of m that comes closest to cutting the number of remaining possible combinations in half.

Implementing this technique on a computer, and looking at what happens in what appears to be the worst case (namely, of the two possibilities for how many combinations remain, suppose we always get the larger), we finish with 1521 tests. This isn't a rigorous proof that it always requires at most 1521 tests, but if it requires more, it is probably not by much.

References

- [1] Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.