Dot Doyle & Dan Teague

The Blood Testing Problem

Suppose that you have a large population 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. In this problem, we are assuming that there are no false positive or false negative tests. 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?

The number of tests might be reduced if the urine could be pooled by putting a number of samples together and then testing the *pooled sample*. Suppose we pool 10 samples together and then test this pooled sample. If the test on the pooled sample is negative, then we know that all 10 individuals in the pooled sample must be negative, and we have checked 10 people with only one test. If, however, the pooled sample tests positive, we know only that at least 1 of the individuals in the sample will test positive. It could be only 1 or all 10 that are positive and we have essentially "wasted" a test (and the money that paid for the test).

The larger the group size for the pooled test, the more we can eliminate with a single test, but the more likely the group is to test positive. Would pooling 5 samples at a time be better than pooling 10 samples at a time? At what point would pooling not be advisable? 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. What is the relationship between the probability of an individual testing positive and the group size that minimizes the total number of tests required?

In this article, we will look at several versions of this classic problem and find solutions using basic algebra, precalculus and data analysis, and calculus. In all versions, we are trying to minimize the expected number of tests performed.

Two Person Problem for Algebra II

The *Mathematics: Modeling Our World* text series has a very nice introduction to this problem in Course 1. Here, the question is the simplest form of the problem. With just two people to test, does it take fewer total tests if the two samples are pooled and tested together first?

Let p represent the probability of a single sample testing positive. So the probability of a single sample testing negative is $1 - p$. To determine the expected number of tests, we need to consider the four possibilities for the two-person sample.

- They are both positive. This happens with probability $P(+\,+) = p^2$.
- They are both negative. This happens with probability $P(-\,-) = (1 - p)^2$.
- The first is positive and the second negative. This happens with probability $P(+\,-) = p(1 - p)$.
- The first is negative and the second positive. This happens with probability $P(-\,+) = (1 - p)p$.

The Expected Number of Tests

Depending on the situation, it could take 1, 2, or 3 tests to determine the sign (+ or –) of the individual samples.

- If both individual samples are negative, then only 1 test of the pooled samples is required.
- If both individual samples are positive, then 3 tests would be required. The initial pooled test would indicate that at least one sample is positive. We would then need to test the first sample separately. This would produce a positive result, which would give us no information about the second sample. The second sample would need to be tested to find that both were positive.
- If the first sample is positive and the second negative, the pooled test would be positive. The first individual test would also be positive and the last test would be negative. So 3 tests would be needed in this case also.
- If the first sample is negative and the second positive, the pooled test would be positive. The first individual test would also be negative. Since the pooled sample was positive and the first individual sample negative, the second sample must be positive. So only 2 tests would be required in this case.

The expected number of tests is $T = 1 \cdot (1 - p)^2 + 2 \cdot (1 - p)p + 3 \cdot p(1 - p) + 3 \cdot p^2$. This simplifies to $T(p) = 1 + 3p - p^2$. We can use this equation to determine when it is to our advantage to pool the samples. Without pooling, exactly 2 tests are always needed. For what probabilities is the expected number of tests from pooling less than 2 (see **Figure 1**)?

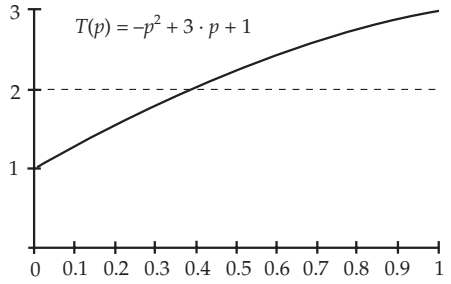


FIGURE 1: EXPECTED NUMBER OF TESTS AS A FUNCTION OF p.

Solving $1 + 3p - p^2 = 2$, we find that pooling should be used only when $p < 0.382$.

One difficulty with this approach is that the analysis is difficult to generalize to more than 2 people. Suppose we had 4 or 8 samples. Considering all of the possible combinations quickly becomes overwhelming. To get around this problem, we can make a simplifying assumption. That is, if the original group tests positive, we will test each of the individual specimens separately. This will give us a "worst case" solution, since, on occasion, we would not need to perform the test on the last member of the group. However, the solution will be much easier to find and will be close to what we would get using the more cumbersome procedure.

Worse Case 2-person Problem

Our worse case analysis assumes that for 2 people we would use either 1 or 3 tests. If both individual samples are negative, then only 1 test of the pooled samples is required. This happens, as before, with probability $(1 - p)^2$. In all other cases (probability $1 - (1 - p)^2$), a total of 3 tests will be used. The expected number of tests will be $t(p) = 1 \cdot (1 - p)^2 + 3 \cdot (1 - (1 - p)^2) = 1 + 4p - 2p^2$ (see **Figure 2**).

With this modification, pooling the two samples would be advantageous whenever $p < 0.293$. We can compare the two expected value functions by considering the difference, $D(p) = t(p) - T(p) = (-2p^2 + 4p + 1) - (-p^2 + 3p + 1) = -p^2 + p$. The difference is a quadratic function whose vertex is (0.5, 0.25). We see that the biggest difference occurs when $p = 0.5$ and at that value the two models differ by only 0.25 tests. Therefore, our worse case solution for the 2-person problem is a very reasonable approximation for the precise method.

In the more advanced formulations of the pooled testing problem, we will use this worse case model, in which, if a pooled test of G individuals is positive, all of the individuals in the pooled sample will be tested separately. So we will either use 1 test or $G + 1$ tests for the G individuals.

Precalculus Solution Using Data Analysis

Precalculus students can handle a more sophisticated (and useful) form of the problem. In this case we will look for a general solution and apply that solution to determine the number of tests needed to find 100 positive individuals in a population of 1,000,000.

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.

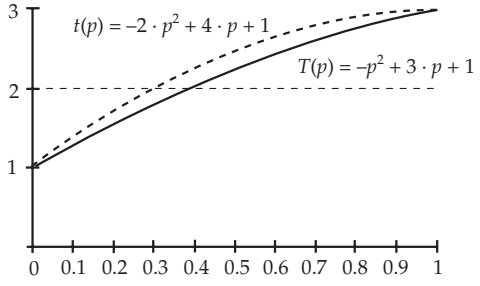


FIGURE 2: COMPARING THE TWO MODELS

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 general pooled testing problem, this is a solution that uses only one pooled 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. In this model we make a second worse case assumption. If a pooled group tests positive, then there is only 1 individual in the group who is positive. Having only 1 person in the group who is positive will result in the maximum number of tests being required. We want to make this maximum number of tests required as small as possible

Expected Number of Tests

Since there are N people to be tested in groups of size G, the initial number of tests needed to test these groups is $\frac{N}{G}$. The probability of an individual testing positive is p, so there are Np people who will test positive. With the worse case assumption that exactly 1 person in each group will test positive, this means that Np of the groups will test positive. Since there are G individuals in each of these groups, there will be NpG people needing to be re-tested. If we do each of these separately, then the number of tests needed for this testing protocol is given by

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

Precalculus students know that the factor N produces a vertical stretch in the graph of T, so the value of G that minimizes the number of tests will not be affected by N. To simplify our work, we can set $N = 1$ for convenience. The

value of G that minimizes $T = \frac{1}{G} + pG$ will also minimize $T = \frac{N}{G} + NpG$.

Using Data Analysis to Find a Function

For a specific value of p, we can determine the group size G that minimizes the number of tests, T, and therefore the costs, by using a graphing calculator. For example, if we let $p = 0.25$, we can graph the function $T(G) = \frac{1}{G} + 0.25G$ and "zoom and trace" or use the "min" key to find that $G = 2$ gives the minimum value of T. If we let $p = 0.01$, we can repeat the process and find the value of G that minimizes $T(G) = \frac{1}{G} + 0.01G$ is $G = 10$. (See **Figure 3**)

By repeating this process for different values of p, we generate **Table 1**:

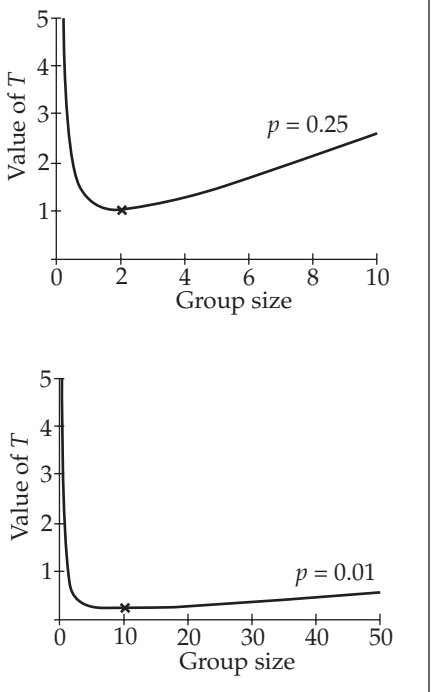


FIGURE 3: GRAPHS OF $T = \frac{1}{G} + pG$ FOR $p = 0.25$ AND $p = 0.01$

The data in this table are ordered pairs that lie on the function describing the relationship between the value of p and the best choice for G. This is the function that will answer our question. 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. The scatterplot is shown in **Figure 4**. Precalculus students should notice that there appears to be both a vertical asymptote at $p = 0$ and a horizontal asymptote at $G = 0$. This suggests some form of a reciprocal function might make a good model, so a power or log-log re-expression would be appropriate.

By re-expressing the data with a log-log plot, we linearize the data (see **Figure 5**).

Since the log-log re-expression linearized the data, we know that the power function model is appropriate. The least-squares line for the re-expressed data is $\ln(G) = 0.00005 - 0.5 \ln(p)$. Solving for G, we find that $G = \frac{1}{\sqrt{p}}$ is the solution to our problem. The group size that will minimize the total number of tests is $G = \frac{1}{\sqrt{p}}$. In our example with 100 individuals testing positive in a population of 1,000,000,

p	0.25	0.20	0.15	0.10	0.05	0.03	0.01	0.005	0.001	0.0005	0.0001
G	2.0	2.24	2.58	3.16	4.47	5.77	10.0	14.14	31.62	44.72	100.0

TABLE 1: BEST G FOR VARIOUS VALUES OF p

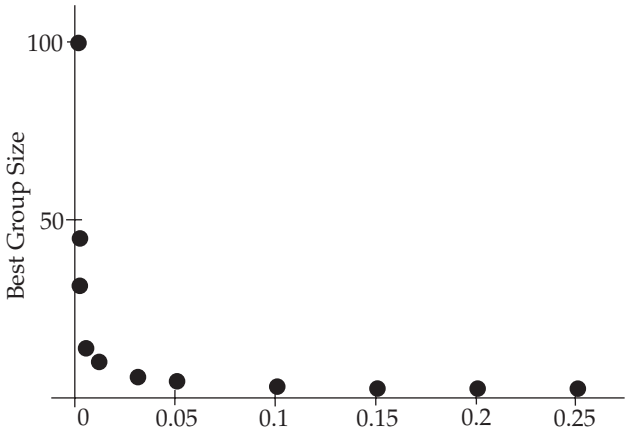


FIGURE 4: SCATTERPLOT OF BEST GROUP SIZE VS. PROBABILITY

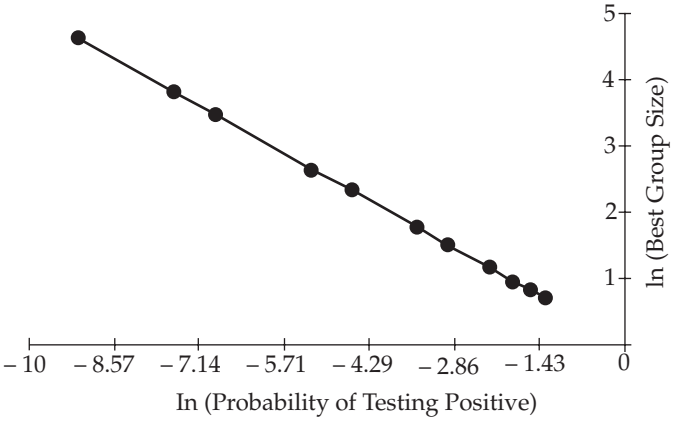


FIGURE 5: LOG-LOG RE-EXPRESSION TO LINEARIZE DATA

$p = 0.0001$, so $G = \frac{1}{\sqrt{0.0001}}$. We should put 100 samples together and test them all together.

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. Remember, this is the worse case. It would never take more than 20,000 tests, and would probably take fewer.

To find the value of p when it would not be useful to pool samples, we solve $2N\sqrt{p} \geq N$, and find that $p \geq \frac{1}{4}$. If $p \geq \frac{1}{4}$, then it takes more total tests when pooling than if we just tested everyone individually initially from the start.

Improving the Solution

One essential aspect of modeling is the importance of taking your first solution and refining and improving it. In our first model, we retested everyone individually. There is no reason to retest everyone individually. We could retest all of the NpG needing to be retested after the first group test in similar groups. We already know that $G = \frac{1}{\sqrt{p}}$ is the optimum group size. 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. To determine the new value of p to use to find G, we need to think carefully about the situation.

There are Np people that we expect to test positive and NpG people remaining to be retested after the first group tests. 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	$\sqrt[8]{p}$
nth Grouping	$\frac{1}{\sqrt[2^n]{p}}$
New Probability	$\sqrt[2^n]{p}$

We know we should stop grouping when the new probability is greater than 0.25. We want to know for what n is $\sqrt[2^n]{p} \geq \frac{1}{4}$. Solving for n, we find that

$$n = \frac{1}{\ln(2)}\ln\left(-\frac{\ln(p)}{\ln(4)}\right).$$

If $p = 0.0001$, then $n \approx 3$, and 3 rounds of pooled tests are needed. How small must p be before 4 rounds are needed?

The model just created works very well, reducing the number of tests dramatically.

In this example of finding 100 positive individuals in a population of 1,000,000, testing in groups of 100, 10, and 3, and then testing everyone remaining individually requires only 10,000 + 1000 + 334 + 300 = 11,634 tests (see **Table 2**). If each test costs $10, then we have saved $9,883,660 over testing individually, and $83,660 over our initial model, which tested in groups only once.

Calculus Solution

In calculus we create the model as in the precalculus version, so the number of tests needed for the testing protocol of 1 group test followed by individual tests is given by

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

Once students have this function, it is a straightforward optimization problem. To determine the best group size G, we differentiate with respect to G, recalling that N and p are parameters. So $\frac{dT}{dG} = -\frac{N}{G^2} + Np$. If $\frac{dT}{dG} = 0$, then $G = \frac{1}{\sqrt{p}}$. What took a lot of effort with data analysis can be done simply with the power of calculus.

From here the solution is the same as that of the extension to the precalculus model. However, our solution 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. However, instead of testing individually, we regrouped and tested in groups, but the model did not

Group Size $\left(\frac{1}{\sqrt{p}}\right)$	Number of Tests $\left(\frac{N}{G}\right)$	Number to Retest (NpG)	New Probability $\sqrt[2^n]{p}$
$\frac{1}{\sqrt{0.0001}} = 1000$	$\frac{1{,}000{,}000}{100} = 10{,}000$	100(100) = 10,000	$\sqrt{0.0001} = 0.01$
$\frac{1}{\sqrt{0.01}} = 10$	$\frac{10{,}000}{10} = 1000$	100(10) = 1000	$\sqrt[4]{0.0001} = 0.1$
$\frac{1}{\sqrt{0.1}} = 3$	$\frac{1000}{3} = 334$	100(3) = 3000	$\sqrt[8]{0.0001} = 0.33 > 0.25$

TABLE 2: FINDING 100 IN 1,000,000 WITH 11,634 TESTS

acknowledge our regrouping. Is it possible to determine the number of tests needed by taking the additional pooled group tests into account from the beginning? A second model extends this initial solution.

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. We test first with groups of size G_1, then group those needing retesting in groups of size G_2. So our initial model allowing for regrouping is

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

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 it possible to rewrite this as a single variable problem? With some encouragement, students will realize that they already know the solution to the last part of the problem (for G_2),

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

because this is the problem of minimizing the number of tests with one test and then testing everyone remaining individually. 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 find that $G_2 = \sqrt{G_1}$. The total number of tests can now be written as a function of the single variable G_1. So,

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

This is now another standard calculus problem. We know that $\frac{dT}{dG_1} = \frac{-N}{G_1^2} + \frac{Np}{G_1^{1/2}}$ and elementary calculus shows that the optimum value for G_1 is $G_1 = p^{-2/3}$. So, if two groupings are used, the sizes of the groups should be $G_1 = p^{-2/3}$ and $G_2 = p^{-1/3}$. Extending the grouping to three continues the pattern. If

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

and we know the solution to the last two group sizes,

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

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

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

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 $G_4 = p^{-1/5}$.

We used 3 regroupings in the earlier solution (see **Table 3**).

Using 3 regroupings in this situation requires only 4000 total tests, and since the latest probability is less than 0.25, we know we could have done even better with 4 regroupings.

Group Size	Number of Tests	Number to Retest	New Probability
$(0.0001)^{-\frac{3}{4}} = 1000$	$\frac{1{,}000{,}000}{1000} = 1000$	100(1000) = 100,000	$\frac{100}{100{,}000} = 0.001$
$(0.0001)^{-\frac{1}{2}} = 100$	$\frac{100{,}000}{100} = 1000$	100(100) = 10,000	$\frac{100}{10{,}000} = 0.01$
$(0.0001)^{-\frac{1}{4}} = 10$	$\frac{10{,}000}{10} = 1000$	100(10) = 10000	$\frac{100}{1000} = 0.1$

TABLE 3: USING 3 REGROUPINGS FROM THE START

The General Solution

From our work above we see that if a total of n groupings are used, the group sizes are given by

$$G_1 = p^{-\frac{n}{n+1}}$$

$$G_2 = p^{\frac{n-1}{n+1}}$$

$$G_3 = p^{\frac{n-2}{n+1}}$$

$$G_n = p^{\frac{1}{n+1}},$$

with the kth group of size $G_k = p^{-\frac{n-(k-1)}{n+1}}$. The total number of tests required with n groupings is

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

This result can be proven by induction, but is generally beyond what most students would be expected to do.

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

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

Solving $\frac{dT}{dn} = 0$ for n we find the optimal number of groupings is $n = -\ln(p) - 1$. For our example, this is $n = -\ln(0.0001) - 1 \approx 8$ groupings of pooled tests. The total number of tests required is given by

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

If $p = 0.0001$ as in our example, 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 2500 tests (and, remember, this is the worse case)!

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

This expression for the group size for the kth regrouping can be simplified. If $G_k = p^{\frac{-\ln(p)-k}{\ln(p)}}$, then $\ln(G_K) = \ln\left(p^{\frac{-\ln(p)-k}{\ln(p)}}\right) = \left(\frac{-\ln(p)-k}{\ln(p)}\right)\ln(p) = -\ln(p) - k$.

Solving again for G_k, we find $G_k = \frac{1}{pe^k}$.

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, since $G_k = \frac{1}{pe^k}$ may be too many specimens to handle in a single group.

Final Comments

This problem offers an important teaching point about mathematical modeling. The importance of iterating the model and refining the solution based on prior work is clear and convincing in this setting. In each approach (using algebra, precalculus, and calculus), we found an initial solution then modified that solution to improve it. This iterative approach is essential to good modeling. In the precalculus solution, we improved the initial solution by violating the assumptions of that solution! We could never do this with a problem in mathematical theory. This creates an interesting discussion about mathematical theory and mathematical practice, and the importance of a good approximate solution over an ideal unrealizable one. Also, the importance of considering "What question does this new solution ask?" is seen in several places. We obtain a solution to one question and immediately use it to answer another. The conversations surrounding the solution and in the process of solving this problem encourage essential aspects of modeling. ❑

REFERENCES

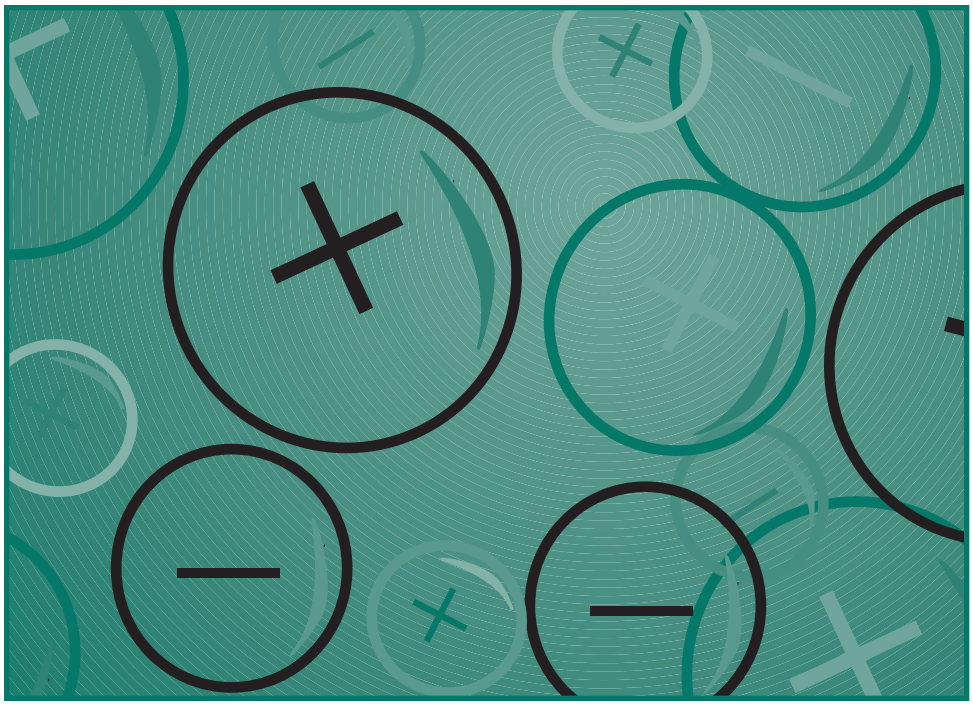
Bartkovich, Kevin, John Goebel, Julie Graves, and Daniel Teague, *Contemporary Calculus through Applications*, Janson Publications, Providence, Rhode Island, 1995.

COMAP, *Mathematics: Modeling Our World*, W. H. Freeman and Company, New York, New York, 1998.

Dilwyn Edwards and Mike Hamson, *Guide of Mathematical Modeling*, CRC Mathematical Guides, CRC Press, Boca Raton, Florida, 1989, pp.199–208.

William Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Ed.* John Wiley and Sons, Inc., New York, 1968, p. 225.

Darian Lauten, and Pat Taylor, "Testing for Steriod Use: When Can You Save Money by Pooling Samples?", *HiMap Pullout Section*, **Consortium**, Number 67, COMAP, Inc. Lexington, Massachusetts, Fall, 1998.



Dan Teague is an instructor of Mathematics at the North Carolina School of Science and Mathematics. He is a Presidential Awardee for North Carolina. You may email him at teague@ncssm.edu or write to: Dan Teague North Carolina School of Science and Mathematics 1219 Broad Street, Durham, NC 27705

Dot Doyle is an Instructor of Mathematics at the North Carolina School of Science and Mathematics. She has served on the editorial panel of NCTM's Mathematics Teacher. You may email her at doyle@ncssm.edu or write to Dot Doyle NCSSM Box 2418 Durham, NC 27715.

Everybody's Problems concerns teaching high school mathematics courses with real-world problems, particularly problems that are suitable for students at all levels.