

# NCAAPMT Calculus Challenge 2010-2011

## Calculus Challenge #6

Solutions due January 19, 2011

### The Group Testing Problem

Suppose that you have a large population ( $N$ ) that you wish to test for a certain characteristic in their urine (all NCAA athletes for steroid use, for example). You will take a sample from all  $N$  individuals and test each sample, with either a positive or a negative result. In this problem we will ignore all false positives and false negatives, since the tests are quite good. Since the number of individuals to be tested is very large, we can expect that the cost of testing will also be large. How can we reduce the number of tests needed to screen everyone and thereby reduce the costs? If the urine samples could be combined or pooled by putting a portion of several samples together and then testing this pooled sample, the number of tests required could be reduced.

Suppose we pool the samples into groups of size  $G$ . That is, we have one test tube (perhaps large) with a portion of the sample from  $G$  individuals in it. 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 to find all positives in a population of size  $N$ ?

Use your model to determine the number of tests needed to find 100 individuals who are positive in a population of 1,000,000.

#### Creating the Model

*One essential aspect in developing any model is to first consider the simplest case that embodies the essence of the problem.* If we cannot solve the simplest version of the problem, then we certainly will not be able to solve a more involved and sophisticated model that is perhaps more realistic. Further, the solution to the simplest situation often is helpful in arriving at a more general solution as we will soon see.

In this setting, the simplest form of the group testing problem is to pool the samples into groups only once. Then test each pooled sample and if it tests negative, remove all the individuals in the group from consideration. For any pooled sample that tests positive, we systematically test every individual in the group to determine if they were positive or negative.

To determine the number of tests needed, we need to consider the two testing settings. First, everyone is put in a group and test the group. Then, anyone in a group that tested positive must be retested individually. The number of tests in total is the sum of the number of tests in the two settings.

- 1) Explain why the number of groups tested is  $\frac{N}{G}$ .

Since there are  $N$  people to be tested, if we put them in groups each of size  $G$ , there will be  $\frac{N}{G}$  groups to test.

- 2) We need to determine many people need to be retested. Recall that the probability of an individual testing positive is  $p$ . Explain why the expected number of tests is modeled by

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

The expected number of people who need retesting can be determined using a probabilistic argument. The probability of testing positive is  $p$ , so the probability of testing negative is  $1 - p$ . The probability that

all  $G$  members of a group test negative is  $(1-p)^G$ , so the probability that at least one person in the group tests positive is  $1-(1-p)^G$ . Therefore,  $1-(1-p)^G$  of the  $\frac{N}{G}$  groups is expected to be retested. Each group contains  $G$  individuals, so the expected number needing retesting is  $(1-(1-p)^G) \cdot \frac{N}{G} \cdot G = N(1-(1-p)^G)$ . We find that the average case model is  $T(G) = N\left(\frac{1}{G} + 1 - (1-p)^G\right)$ .

3) Given the model,  $T(G) = \frac{N}{G} + (1-(1-p)^G)N$ , explain why the techniques of calculus are not suitable for finding the value of  $G$  that minimizes  $T$ .

### Linear Approximations

In practice, we often replace a messy or difficult expression with its linear approximation. In this case, we want to replace the expression  $(1-(1-p)^G)$  in the equation above with its linear approximation. We will then have a function for which we can indeed find a solution using calculus.

4) a) Find the linear approximation for  $f(G) = (1-(1-p)^G)$  as a function of  $G$  at  $G = 0$ .

If we consider  $f(G) = 1 - (1-p)^G$ , then the tangent line approximation at  $G = 0$  is  $A(G) = -\ln(1-p)G$ . This substitution creates a tractable one-group solution,  $T(G) = N\left(\frac{1}{G} - \ln(1-p)G\right)$ , but the algebra is very cumbersome.

b) Find the linear approximation for  $f(p) = (1-(1-p)^G)$  as a function of  $p$  at  $p = 0$ .

If we consider  $f(p) = 1 - (1-p)^G$ , then  $1 - (1-p)^G \approx pG$  close to  $p = 0$ .

This is equivalent to a worst case analysis. The probability of testing positive is  $p$ , so we expect that  $Np$  people will test positive. The worst case assumption (if a group tests positive, exactly one person in that group will test positive) means that  $Np$  of the groups will test positive. Consequently, there are  $NpG$  people requiring retesting. The number of tests needed for this testing protocol can be described by the equation  $T = \frac{N}{G} + NpG$ .

5) Substitute the simpler approximation from 4) into the function  $T(G) = \frac{N}{G} + N(1-(1-p)^G)$  and find the value of  $G$  that minimizes  $T$ .

Using  $T = \frac{N}{G} + NpG$ , to determine the best group size  $G$ , we differentiate with respect to  $G$ . So,  $\frac{dT}{dG} = -\frac{N}{G^2} + Np$ . If  $\frac{dT}{dG} = 0$ , then  $G = \frac{1}{\sqrt{p}}$  and the total number of tests needed is  $T = 2N\sqrt{p}$ .

6) Use the solution from 5) to determine the number of tests needed if  $N = 1,000,000$  and  $p = 0.0001$ . What group sizes are required and how many total tests are needed?

If  $N = 1,000,000$  and  $p = 0.0001$ , then we are looking for 100 needles in this 1,000,000 straw haystack. We will need only  $T = 2(1,000,000)\sqrt{0.0001} = 20,000$  tests. This is the worst case, so it will never take more than this.

7) Use your solution to 5) to determine which values of  $p$  would testing in groups require more total tests than just testing all  $N$  individuals separately? For example, if  $p = 0.75$ , it would be foolish to test in groups.

If the total number of tests from grouping is greater than  $N$ , then it would be foolish to test in groups. So, if  $2N\sqrt{p} \geq N$ , then just test everyone individually. Solving for  $p$ , we find Further, if  $p \geq \frac{1}{4}$  it is counterproductive to test in groups.

### Improving the Model

In developing the model, we have assumed that we would test only once in groups and then test everyone remaining individually. Clearly, we could regroup those remaining and do a second round of group tests, before testing everyone remaining individually.

8) Use the relationship between  $G$  and  $p$  from 5) that minimizes the number of tests required and find the group size in terms of  $p$  for a second round of group tests. That is, instead of testing all members of groups that tested positive in the first test, we form new groups and retest the groups. What group size would you use in the second group tests? Why is the value of  $p$  different in the second grouping than in the first?

In reality, 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	$\sqrt[8]{p}$
	$\vdots$		$\vdots$

$$\text{nth Grouping} \quad \frac{1}{\sqrt[n]{p}} \quad \text{New Probability} \quad \sqrt[n]{p}$$

When do you stop grouping and test everyone individually? We want to know the smallest value of  $n$  for which  $\sqrt[n]{p} \geq \frac{1}{4}$ . Solving for  $n$ , we find that

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

The number of iterations required is quite small. If  $p > 0.000000001$ , then  $n < 4$ .

9) If you repeatedly used the solution to 5) and the information gained in 8) in retesting, determine the total number of tests required to find 100 individuals who are positive in a population of 1,000,000 by filling in the table below. Remember to use your solution to 7) as a stopping criterion.

Round	Number to be Tested or Retested	Probability of Testing Positive	Size of Optimal Group	Number of Group Tests
1	1,000,000	0.0001	$\frac{1}{\sqrt{0.0001}} = 100$	10,000
2	100 groups of 100 10,000	$\frac{100}{(100)(100)} = 0.01$	$\frac{1}{\sqrt{0.01}} = 10$	10,000 + 1,000
3	100 groups of 10 1,000	$\frac{100}{(100)(10)} = 0.10$	$\frac{1}{\sqrt{0.1}} = 3$	10,000 + 1,000 + 334
4	100 groups of 3 300	$\frac{100}{(100)(3)} = 0.33$	Since $p = 0.33$ , we should test individually	10,000 + 1,000 + 334 + 300
5				
6				

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 at worst **11,634** tests.

### Only if you are interested:

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.

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 rounds of group tests before individual testing. So,

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

We see 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 in terms of a single variable model?

We already know the solution to the last part of the problem,

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

which is the problem of minimizing the number of tests with one test round of group tests and then retesting everyone remaining individually. Recall that this group size was independent of the number being tested. So, in fact, we know that  $G_2 = \frac{1}{\sqrt{p^*}}$ , 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  $G_1$ :

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

See what you can make out of this modification for 2-groups, 3-groups, ...  $k$ -groups. Write to me if you want to see how the rest of this works out. We can find 100 out of 1,000,000 in at most 2,500 tests.