Group Testing: Single Pooling and Multiple Pooling

Michiel Smid[*]

October 29, 2020

1 Introduction

Consider a group of people, some of which are infected by a virus such as COVID-19 or Ebola. We would like to identify the infected people in this group. Assume that we have a *sample* for each person in the group. For example, when testing for COVID-19, the sample is collected through a nasal swab, whereas a blood sample is used when testing for Ebola. These samples are analyzed in a laboratory. The obvious approach is to analyze each sample individually. The drawback is that one laboratory test is needed for each person. In 1943, Robert Dorfman[1] proposed *single pooling*: For a given integer s, we divide the people into subgroups, each having size s. For each subgroup, we combine (parts of) the samples of the people in this subgroup and do one laboratory test for the resulting mix. If the test is *negative*, then none of the people in the subgroup is infected. Otherwise, i.e., if the test is *positive*, we know that at least one person in the subgroup is infected. In the latter case, we use the remainders of the samples to test each person in the subgroup individually. Assuming that we have a good estimate for the number of infected people, Dorfman showed that, by choosing the *pool size* s appropriately, the number of laboratory tests can be reduced significantly. In these notes, we will work out the details of Dorfman's approach. We will also present the generalization to *multiple pooling*, which was proposed by Broder and Kumar[2] in 2020.

2 Notation and Assumptions

1. The number of people in the group is denoted by n. We number these people, arbitrarily, as $P_1, P_2, \ldots, P_n$.

[*] School of Computer Science, Carleton University, Ottawa, Canada.

[1] The Detection of Defective Members of Large Populations, The Annals of Mathematical Statistics, **14(4)**, pages 436–440.

[2] A Note on Double Pooling Tests, `https://arxiv.org/abs/2004.01684`.

2. The number of infected people in the group is denoted by k. We assume that we know the exact value of k. (Even though this is not a realistic assumption, a reasonable estimate of k can be obtained from previous testing results.)

3. We have a procedure $\textsc{Test}(T)$, which takes as input a non-empty subset T of $\{1, 2, \ldots, n\}$. This procedure combines (parts of) the samples of the people in the subset $\{P_i : i \in T\}$ and tests the mix. The procedure returns either

   - *negative*, meaning that none of the people in $\{P_i : i \in T\}$ is infected, or
   - *positive*, meaning that at least one person in $\{P_i : i \in T\}$ is infected. In this case, if $|T| \geq 2$, we do not know who is/are infected.

   (For COVID-19 testing, this is possible for sets T of sizes up to 64. See the references in the paper by Broder and Kumar.)

4. We assume that the output of the procedure $\textsc{Test}(T)$ is always correct. Thus, there are no *false positives* and no *false negatives*. (This is not a realistic assumption, but it makes the analysis much simpler.)

Our goal is to design a *testing algorithm*, i.e., a sequence of calls to the procedure $\textsc{Test}(T)$, using different sets T, such that, at termination, we know for each i in $\{1, 2, \ldots, n\}$, whether or not person P_i is infected. The goal is to minimize the number of calls to the procedure $\textsc{Test}$.

An obvious algorithm tests each sample *individually*: For each $i = 1, 2, \ldots, n$, we make one call to $\textsc{Test}(\{i\})$. In this way, the procedure is called n times.

In the rest of these notes, we will show that, if the number of infected people is "small", we can identify them using a much smaller number of calls to the procedure $\textsc{Test}$.

3 Single Pooling

As was mentioned in Section 1, single pooling was proposed by Robert Dorfman in 1943. It is being used for COVID-19 testing in Germany and Israel; see the references in the paper by Broder and Kumar.

Recall that n denotes the number of people to be tested. In single pooling, we choose an integer $s \geq 2$ such that n is a multiple of s. Consider a permutation Π of the people $P_1, P_2, \ldots, P_n$. We divide this permutation into n/s *blocks*: The first s people in Π form the first block, the next s people form the second block, etc. For each block, we run the procedure $\textsc{Test}$ on the indices of the people in this block. If the result for a block is *negative*, then we know that none of the people in this block is infected. On the other hand, if the result for a block is *positive*, then at least one person in this block is infected. In the latter case, we test the sample of each person in this block individually.

Observe that one person may be involved in two calls to the procedure $\textsc{Test}$. Therefore, we divide each person's sample into two subsamples. In each call to $\textsc{Test}$, we use a different subsample.

It is clear that this testing algorithm is correct: For each i with $1 \leq i \leq n$, it correctly determines whether or not P_i is infected. The number of calls to the procedure TEST depends on the permutation Π. Since, at the start of the algorithm, we do not know who is infected and who is not infected, we choose Π uniformly at random from the set of all $n!$ permutations. Here is a formal description of the algorithm:

```
Algorithm SinglePooling($n, s$)

Comment: $n$ denotes the number of people and $s$ denotes the block size. We assume
that $n$ is a multiple of $s$. The people to be tested are denoted by $P_1, P_2, \ldots, P_n$.

Let $\Pi$ be a uniformly random permutation of $P_1, P_2, \ldots, P_n$.
Divide $\Pi$ into $n/s$ blocks, each of size $s$.
For each $j = 1, 2, \ldots, n/s$, do the following:
    1. Run the procedure Test on the indices of the people in the $j$-th block of $\Pi$.
    2. If the result is negative, then none of the people in this block is infected.
    3. If the result is positive, then run the procedure Test on the index of each person
       in this block individually.
```

Consider the random variable

$$X = \text{the total number of calls to the procedure Test when running algorithm SinglePooling}(n, s).$$

We are going to determine the expected value $\mathbb{E}(X)$ of X.

It is natural to introduce the indicator random variables $X_1, X_2, \ldots, X_n$, where

$$X_i = \begin{cases} 1 & \text{if the sample of } P_i \text{ is tested individually,} \\ 0 & \text{otherwise.} \end{cases}$$

Let I be the set of indices of the infected people, and let N be the set of indices of the non-infected people. Observe that

$$X = \frac{n}{s} + \sum_{i=1}^{n} X_i = \frac{n}{s} + \sum_{i \in I} X_i + \sum_{i \in N} X_i. \tag{1}$$

Recall that k denotes the number of infected people, i.e., $k = |I|$. We assume that $2 \leq s \leq n - k$.

If i is an index in I, i.e., P_i is infected, then $X_i = 1$, because, no matter which permutation Π is chosen, the sample of P_i is tested individually. Thus,

$$\text{for each index } i \text{ in } I, \ \mathbb{E}(X_i) = 1. \tag{2}$$

Let i be an index in N. Thus, P_i is not infected. We are going to determine the expected value $\mathbb{E}(X_i)$ of the random variable X_i.

Observe that $X_i = 1$ if and only if at least one person in P_i's block in the permutation Π is infected. Consider the event

$$A = \text{“none of the other } s-1 \text{ people in } P_i\text{'s block in } \Pi \text{ is infected”.}$$

Then

$$\mathbb{E}(X_i) = \Pr(X_i = 1) = 1 - \Pr(A).$$

To determine $\Pr(A)$, we make the following observations:

- For a given permutation Π, whether or not the event A occurs is completely determined by (i) the position of P_i in Π and (ii) the positions in Π of the k infected people.
- In any permutation Π, person P_i is at one of n possible positions. The k infected people can be in any of the remaining $n-1$ positions. Thus, there are $\binom{n-1}{k}$ many possible subsets for these k positions.
- As above, in any permutation Π, P_i is at one of n possible positions. The event A occurs if and only if all k infected people are outside of P_i's block in the permutation. Thus, these infected people can be in any of $n-s$ positions. There are $\binom{n-s}{k}$ many possible subsets for these positions.

It follows that

$$\Pr(A) = \frac{n\binom{n-s}{k}}{n\binom{n-1}{k}} = \frac{\binom{n-s}{k}}{\binom{n-1}{k}}. \tag{3}$$

In case you do not like the above argument, here is an alternative derivation. We are going to count the number of permutations Π for which the event A occurs. By dividing this number by $n!$, we obtain $\Pr(A)$. In the following four steps, we specify a unique permutation for which A occurs.

1. Choose a position for P_i in the permutation. There are n ways to do this. The chosen position determines the block in the final permutation that contains P_i. Below, we denote this block by B.
2. Choose a subset of size $s-1$ from the set of $n-k-1$ non-infected people that are not equal to P_i. There are $\binom{n-k-1}{s-1}$ ways to do this.
3. Choose a permutation of the people in the subset that was chosen in 2. Add these people, according to the chosen permutation, to the block B. There are $(s-1)!$ ways to do this.
4. Choose a permutation of the remaining $n-s$ people, and add them, according to the chosen permutation, to the positions outside of the block B. There are $(n-s)!$ ways to do this.

By the Product Rule, the total number of permutations for which the event A occurs is equal to

$$n \cdot \binom{n-k-1}{s-1} \cdot (s-1)! \cdot (n-s)!.$$

Therefore,

$$\Pr(A) = \frac{n \cdot \binom{n-k-1}{s-1} \cdot (s-1)! \cdot (n-s)!}{n!},$$

which, by basic algebra, is equal to the fraction in (3).

We have shown that

$$\text{for each index } i \text{ in } N,\ \mathbb{E}\left(X_i\right) = 1 - \Pr(A) = 1 - \frac{\binom{n-s}{k}}{\binom{n-1}{k}}. \tag{4}$$

By applying the Linearity of Expectation to (1), substituting (2) and (4), and using the facts that $|I| = k$ and $|N| = n-k$, we obtain

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}\left(\frac{n}{s} + \sum_{i \in I} X_i + \sum_{i \in N} X_i\right) \\
&= \frac{n}{s} + \sum_{i \in I} \mathbb{E}\left(X_i\right) + \sum_{i \in N} \mathbb{E}\left(X_i\right) \\
&= \frac{n}{s} + \sum_{i \in I} 1 + \sum_{i \in N} \left(1 - \frac{\binom{n-s}{k}}{\binom{n-1}{k}}\right) \\
&= \frac{n}{s} + k + (n-k)\left(1 - \frac{\binom{n-s}{k}}{\binom{n-1}{k}}\right) \\
&= n + \frac{n}{s} - (n-k)\frac{\binom{n-s}{k}}{\binom{n-1}{k}}. \qquad (5)
\end{aligned}$$

Recall that our goal is to minimize the number of calls to the procedure TEST. Thus, for fixed values of n and k, we want to choose s such that $\mathbb{E}(X)$ is minimum.

Remark: The analysis in Dorfman's paper, as well as in the paper by Broder and Kumar, is different from the one given above. They consider a "streaming" version of the problem: People arrive one by one. The current person to arrive has a probability p of being infected, independently of all other people. In our notation, this corresponds to the case when $n \to \infty$ and $k/n \to p$.

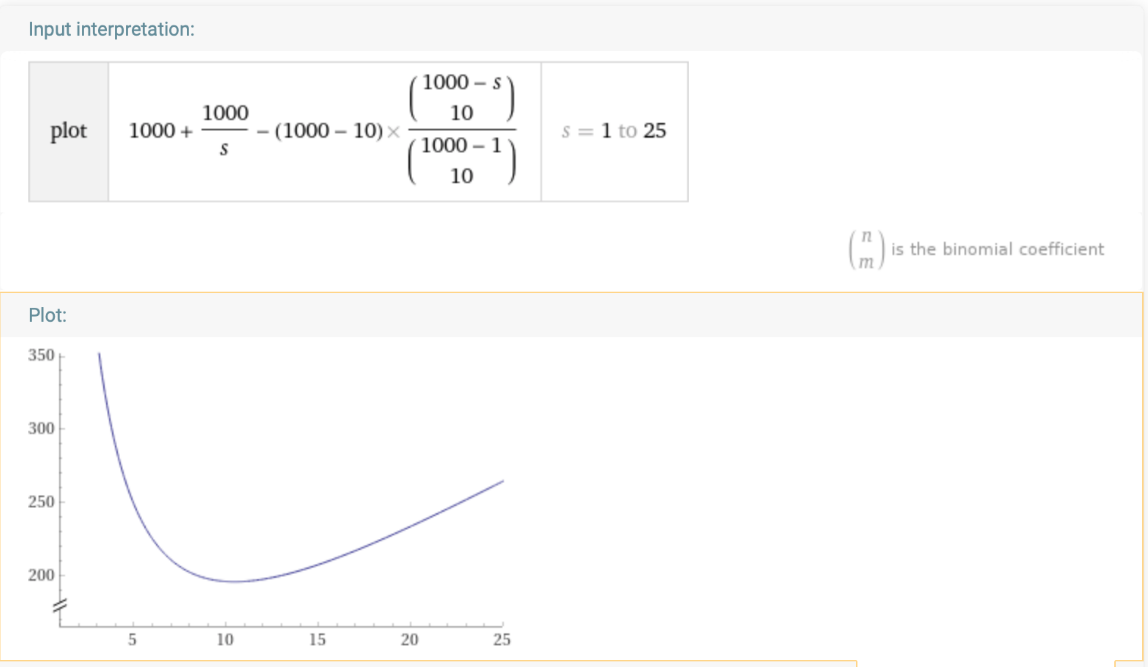
3.1 An Example

We consider the case when $n = 1000$ and $k = 10$. We use the expression for $\mathbb{E}(X)$ as given in (5). Using Wolfram Alpha, we get:

- For $s = 5$, $\mathbb{E}(X) \approx 249$.

- For $s = 10$, $\mathbb{E}(X) \approx 196$.
- For $s = 20$, $\mathbb{E}(X) \approx 234$.

Wolfram Alpha gives us the graph of $\mathbb{E}(X)$ as a function of s; see the figure below. We see that the minimum occurs around $s = 10$. Thus, for this example, testing the sample of each person individually requires 1000 calls to the procedure TEST. By using single pooling with $s = 10$, the expected number of calls is only 196.



Input interpretation:

plot	$1000 + \frac{1000}{s} - (1000 - 10) \times \frac{\binom{1000-s}{10}}{\binom{1000-1}{10}}$	$s = 1$ to 25

$\binom{n}{m}$ is the binomial coefficient

Plot:

4 Multiple Pooling

As was mentioned in Section 1, multiple pooling was proposed by Broder and Kumar in 2020. It is based on the *power of two choices* paradigm.

Let $c \geq 1$ be an integer. Instead of using one permutation Π, as we did in algorithm SINGLEPOOLING, we use c permutations. For each $\ell = 1, 2, \ldots, c$, we run the procedure TEST on each of the n/s blocks of the ℓ-th permutation.

Consider a person P_i. If there is an ℓ for which the result of P_i's block in the ℓ-th permutation is *negative*, then we know that P_i is not infected. Otherwise, P_i may be infected. In the latter case, we test the sample of P_i individually.

Observe that one person may be involved in $c+1$ calls to the procedure TEST. Therefore, we divide each person's sample into $c+1$ subsamples. In each call to TEST, we use a different subsample. Here is a formal description of the algorithm:

Algorithm MULTIPLEPOOLING(n, c, s)

Comment: n denotes the number of people, c denotes the number of permutations, and s denotes the block size for each permutation. We assume that n is a multiple of s. The people to be tested are denoted by $P_1, P_2, \dots, P_n$.

Let $\Pi_1, \Pi_2, \dots, \Pi_c$ be uniformly random permutations of $P_1, P_2, \dots, P_n$. These permutations are chosen independently of each other.
For each $\ell = 1, 2, \dots, c$, do the following:

1. Divide Π_ℓ into n/s blocks, each of size s.
2. For each $j = 1, 2, \dots, n/s$, run the procedure TEST on the indices of the people in the j-th block of Π_ℓ.

For each $i = 1, 2, \dots, n$, do the following:

1. If there is an ℓ for which the result of P_i's block in the permutation Π_ℓ is *negative*, then P_i is not infected.
2. Otherwise, run the procedure TEST$(\{i\})$, i.e., run the procedure TEST on the index i of P_i.

It is not difficult to verify that this testing algorithm is correct. That is, for each i with $1 \leq i \leq n$, it correctly determines whether or not P_i is infected.

Consider the random variables

$$Y = \text{the total number of calls to the procedure TEST when running algorithm MULTIPLEPOOLING}(n, c, s)$$

and, for $i = 1, 2, \dots, n$,

$$Y_i = \begin{cases} 1 & \text{if the sample of } P_i \text{ is tested individually,} \\ 0 & \text{otherwise.} \end{cases}$$

As before, let I be the set of indices of the infected people, and let N be the set of indices of the non-infected people. Recall that $k = |I|$. We assume that $2 \leq s \leq n - k$. Observe that

$$Y = \frac{cn}{s} + \sum_{i \in I} Y_i + \sum_{i \in N} Y_i.$$

If i is an index in I, then $Y_i = 1$, no matter which permutations are chosen. Thus, in this case, $\mathbb{E}(Y_i) = 1$.

Let i be an index in N. For each $\ell = 1, 2, \dots, c$, consider the event

$$A_\ell = \text{“none of the other } s-1 \text{ people in } P_i\text{'s block in } \Pi_\ell \text{ is infected”.}$$

Then $Y_i = 1$ if and only if the event

$$\overline{A}_1 \wedge \overline{A}_2 \wedge \cdots \wedge \overline{A}_c$$

occurs, where $\overline{A}_\ell$ denotes the complement of A_ℓ. Since the c permutations are chosen independently, the events $\overline{A}_\ell$, $\ell = 1, 2, \ldots, c$, are mutually independent. Therefore, if we let

$$p = \frac{\binom{n-s}{k}}{\binom{n-1}{k}},$$

then it follows from (3) that

$$\mathbb{E}\left(Y_i\right) = \Pr\left(Y_i = 1\right) = \Pr\left(\bigwedge_{\ell=1}^{c} \overline{A}_\ell\right) = \prod_{\ell=1}^{c} \Pr\left(\overline{A}_\ell\right) = \prod_{\ell=1}^{c}(1-p) = (1-p)^c.$$

By putting everything together, we conclude that

$$\begin{aligned} \mathbb{E}(Y) &= \frac{cn}{s} + k + (n-k)(1-p)^c \\ &= \frac{cn}{s} + k + (n-k)\left(1 - \frac{\binom{n-s}{k}}{\binom{n-1}{k}}\right)^c . \end{aligned} \tag{6}$$

4.1 An Example

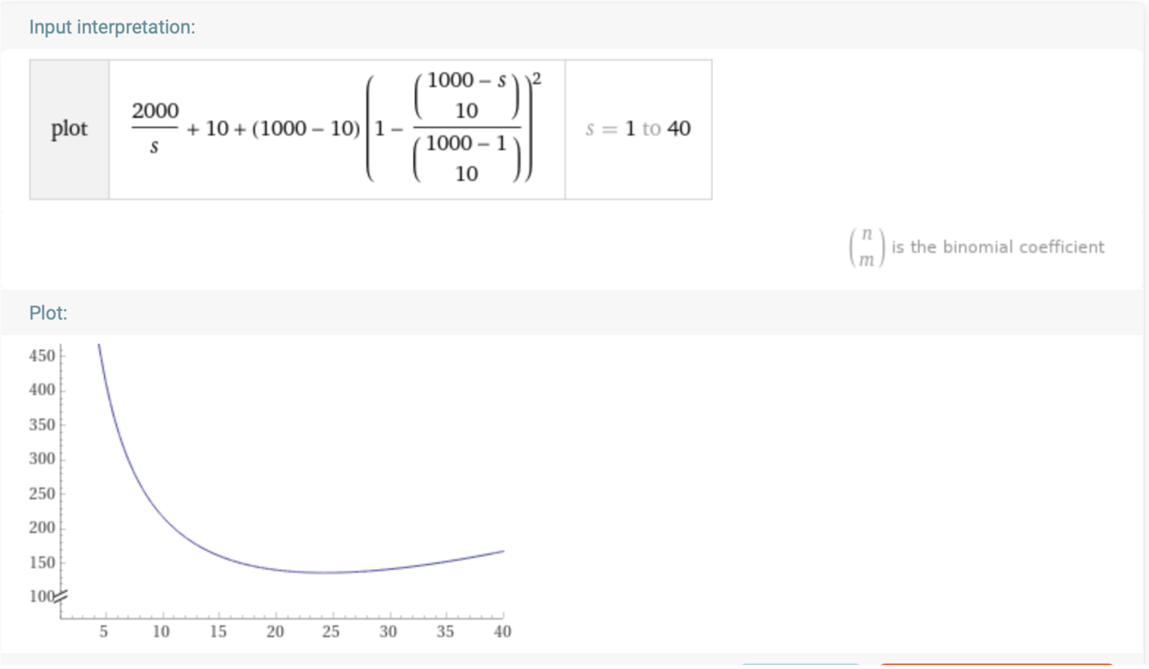
As in Section 3.1, we consider the case when $n = 1000$ and $k = 10$. We use the expression for $\mathbb{E}(Y)$ as given in (6). For each $c \in \{2, 3, 4\}$, we will determine the value of s for which $\mathbb{E}(Y)$ is minimum. All figures below were obtained using Wolfram Alpha.

4.1.1 Using Two Permutations

For $c = 2$, Wolfram Alpha gives the following values:

- For $s = 5$, $\mathbb{E}(Y) \approx 412$.
- For $s = 10$, $\mathbb{E}(Y) \approx 218$.
- For $s = 20$, $\mathbb{E}(Y) \approx 141$.
- For $s = 25$, $\mathbb{E}(Y) \approx 137$.
- For $s = 30$, $\mathbb{E}(Y) \approx 142$. (Note that n is not a multiple of s.)

Below, you see the graph of $\mathbb{E}(Y)$ as a function of s. We see that the minimum occurs around $s = 25$.



Input interpretation:

plot	$\frac{2000}{s} + 10 + (1000-10)\left(1 - \frac{\binom{1000-s}{10}}{\binom{1000-1}{10}}\right)^2$	$s = 1$ to 40

$\binom{n}{m}$ is the binomial coefficient

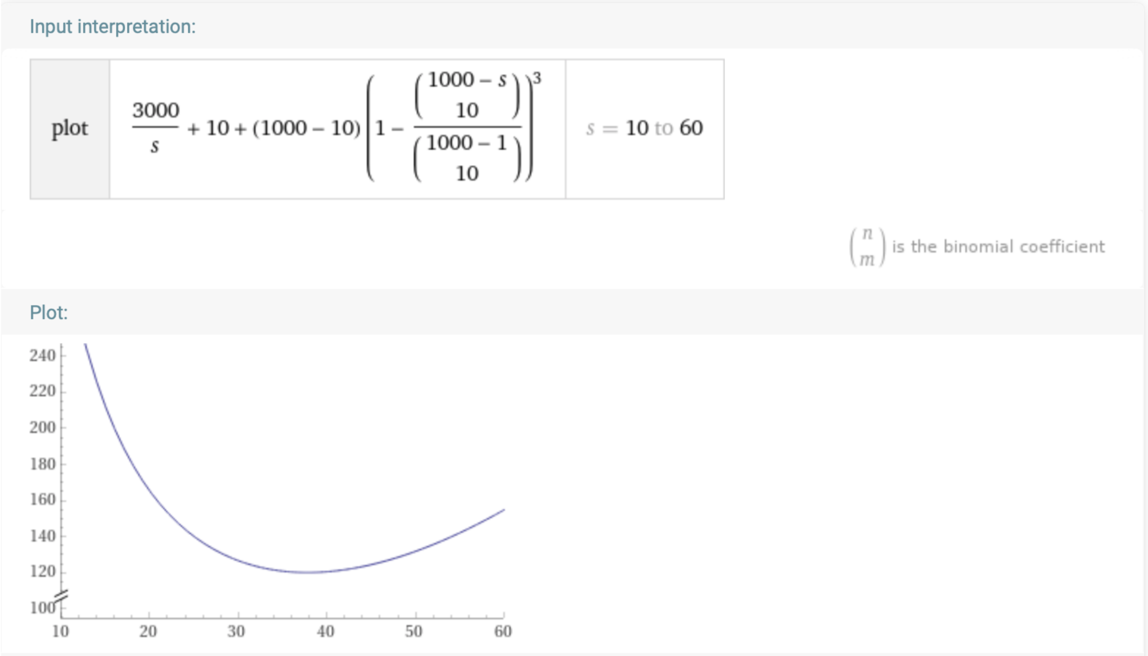
Plot:

4.1.2 Using Three Permutations

For $c = 3$, Wolfram Alpha gives the following value:

- For $s = 38$, $\mathbb{E}(Y) \approx 120$. (Note that n is not a multiple of s.)

Below, you see the graph of $\mathbb{E}(Y)$ as a function of s. We see that the minimum occurs around $s = 38$.



Input interpretation:

plot	$\frac{3000}{s} + 10 + (1000-10)\left(1 - \frac{\binom{1000-s}{10}}{\binom{1000-1}{10}}\right)^3$	$s = 10$ to 60

$\binom{n}{m}$ is the binomial coefficient

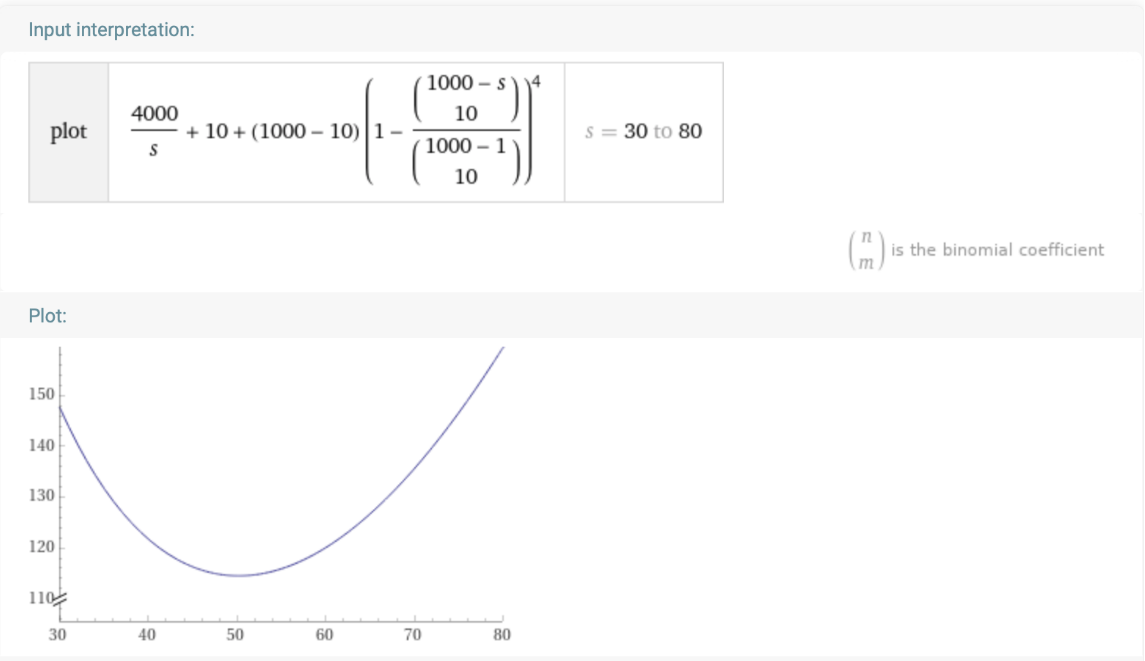
Plot:

4.1.3 Using Four Permutations

For $c = 4$, Wolfram Alpha gives the following value:

- For $s = 50$, $\mathbb{E}(Y) \approx 114.5$.

Below, you see the graph of $\mathbb{E}(Y)$ as a function of s. We see that the minimum occurs around $s = 50$.



Input interpretation:

plot	$\frac{4000}{s} + 10 + (1000 - 10)\left(1 - \frac{\binom{1000-s}{10}}{\binom{1000-1}{10}}\right)^4$	$s = 30$ to 80

$\binom{n}{m}$ is the binomial coefficient

Plot:

4.1.4 Conclusion

For the case when $n = 1000$ and $k = 10$, we have obtained the following results:

- Testing the sample of each person individually requires 1000 calls to the procedure TEST.

- By using single pooling with $s = 10$, the expected number of calls to TEST is about 196.

- By using multiple pooling with $c = 2$ and $s = 25$, the expected number of calls to TEST is about 137.

- By using multiple pooling with $c = 3$ and $s = 38$, the expected number of calls to TEST is about 120.

- By using multiple pooling with $c = 4$ and $s = 50$, the expected number of calls to TEST is about 114.5.

5 Further Reading

Wikipedia has an article on Group Testing:

https://en.wikipedia.org/wiki/Group_testing

For a list of countries that use pooling for COVID-19 testing, go to

https://en.wikipedia.org/wiki/List_of_countries_implementing_pool_testing_strategy_against_COVID-19