On the Number of Tests Needed for the Pooled Testing Halving Scheme

Sheldon M. Ross

Department of Industrial and Systems Engineering

University of Southern California

Los Angeles, CA 90089

smross@usc.edu

This material is based on work supported by, or in part, by the National Science Foundation under contract/grant CMMI1662442.

Abstract

We develop recursions for computing the mean and variance of the number of pooled tests needed by the halving scheme to determine the disease positive members of a population. It is assumed that each population member is independently positive with probability p, that the individual blood samples can be pooled to test whether or not at least one member of that pooled group is positive, and that a positive tested group is then split in half and the process continued.

1 Introduction

Suppose that we want to determine which members of a population have a certain disease, and that whether or not any of the members of a group has the disease can be ascertained by a single pooled test. Assuming that each member of the population independently has the disease with known probability p, the objective is to find a pooling scheme that minimizes the expected number of tests to classify all members of the population as being disease positive or negative.

The *halving scheme* for pooled testing was introduced in [**1**]. Say that a group is positive if it is known that at least one of its pooled members has the disease. In the halving scheme one breaks up the population into sets of size n and does a pooled test on a set. If it comes up positive, then the set is randomly split into two new groups of equal size (or as near as possible) and each of these groups is given a pooled test. This continues until the disease state of all members of the population has been determined. (Actually, if a positive group is broken in two subgroups, and the first one of its subgroups tests negative, then the other subgroup must be positive and so need not be tested as a whole but should just be randomly split.) Let X_n denote the number of tests needed to classify a group of size n when you start by testing all n as a pooled group and then use the halving procedure afterwards. Recursions for the probability mass function of X_n when $n = 15$ were given in [**1**] by an approach that does not appear to scale up for larger values of n. In the recent paper [**2**], the halving scheme was proposed as a possibility for use in the Covid-19 pandemic in places where test kits are limited. It is suggested in [**2**] that to minimize $E[X_n]/n$, the average number of tests per individual, n be chosen as close as possible to $n_0 \equiv -\log(2)/\log(1-p)$, which would make $(1-p)^{n_0}$, the probability of a positive result, approximately equal .5. The mean number of tests required per person was determined by a simulation in [**2**], and an analytic approximation was provided. Whereas [**1**] and [**2**] assume that p is know, the paper [**3**] deals with the problem of estimating p when a pooled scheme is used.

In this paper, we show how to derive $E[X_n]$ and $\text{Var}(X_n)$ by easily computed recursions. In doing so, we also indicate that the initial choice of pooling n_0, while not optimal is quite close. Indeed, letting $n(\text{opt})$ be the value of n that minimizes $\frac{E[X(n)]}{n}$; if $p = .05$ then $n_0 = 13.5$ whereas $n(\text{opt}) = 13$; if $p = .01$ then $n_0 = 68.9$ whereas $n(\text{opt}) = 75$; if $p = .001$ then $n_0 = 692.8$ whereas $n(\text{opt}) = 683$; and if $p = .0001$ then $n_0 = 6931.13$ whereas $n(\text{opt}) = 6827$.

In our analysis we will suppose that if the pooled test of an odd number of people comes up positive, then the first of its two subgroups to be tested is the smaller one. In Section 2 we derive a recursion for $E[X_n]$. In Section 3 we give the R code for obtaining $\text{argmin}_n E[X(n)]/n$ and $\min E[X_n]/n$, and comment on what can be done when p is initially unknown. Along the way, we obtain counterexamples showing that (a) the halving scheme of splitting a positive group into two equal subgroups need not be optimal, and (b) $-E[X_n]/n$ is not necessarily a unimodal function of n (e.g., it is not necessarily true that $E[X_n]/n$ first decreases and then increases). Finally, in Section 4 we derive a recursion for computing $\mathrm{Var}(X_n)$.

2 Recursive Formula for $E[X_n]$

Let $q = 1 - p$. Let I be the indicator of the event that the group of size n tests positive. That is, $I = 1$ if the group tests positive, and 0 otherwise. Then, with Y_n denoting the number of tests needed to identify all members of a group of size n known to be positive,

$$E[X_n|I] \quad = \quad 1 + I\,E[Y_n]$$

Using that $P(I = 1) = 1 - q^n$, we see that

$$E[X_n] \quad = \quad 1 + (1 - q^n)E[Y_n] \tag{1}$$

To determine a recursion for $E[Y_n]$, let J be the indicator function of the event that the first one of the 2 splittings of the positive group of size n (e.g., the one of size $\frac{n-1}{2}$ when n is odd) tests positive. Then

$$P(J = 1) = \begin{cases} \frac{1-q^{n/2}}{1-q^n} = \frac{1}{1+q^{n/2}} & \text{if } n \text{ is even} \\ \frac{1-q^{(n-1)/2}}{1-q^n} & \text{if } n \text{ is odd} \end{cases}$$

Because the second subgroup will be known to be positive if the first tests negative, we see that when n is an even positive integer

$$\begin{aligned} E[Y_n|J] &= 1 + E[Y_{\frac{n}{2}}] + JE[X_{\frac{n}{2}}] && (2) \\ &= 1 + E[Y_{\frac{n}{2}}] + J(1 + (1 - q^{n/2})E[Y_{\frac{n}{2}}]) && (3) \end{aligned}$$

Thus, when n is even

$$\begin{aligned} E[Y_n] &= 1 + E[Y_{\frac{n}{2}}] + \frac{1 + (1 - q^{n/2})E[Y_{\frac{n}{2}}]}{1 + q^{n/2}} \\ &= \frac{2 + q^{n/2} + 2E[Y_{\frac{n}{2}}]}{1 + q^{n/2}} && (4) \end{aligned}$$

Now $Y(1) = 0$, and when $n \geqslant 3$ is odd

$$\begin{aligned} E[Y_n|J] &= 1 + J\left(E[Y_{\frac{n-1}{2}}] + E[X_{\frac{n+1}{2}}]\right) + (1 - J)E[Y_{\frac{n+1}{2}}] && (5) \\ &= 1 + J\left(E[Y_{\frac{n-1}{2}}] + 1 - q^{(n+1)/2}E[Y_{\frac{n+1}{2}}]\right) + E[Y_{\frac{n+1}{2}}] \end{aligned}$$

Thus, when $n \geqslant 3$ is odd

$$\begin{aligned} E[Y_n] &= 1 + E[Y_{\frac{n+1}{2}}] + \frac{1 - q^{(n-1)/2}}{1 - q^n}\left(E[Y_{\frac{n-1}{2}}] + 1 - q^{(n+1)/2}E[Y_{\frac{n+1}{2}}]\right) \\ &= 1 + \frac{1 - q^{(n-1)/2}}{1 - q^n}(E[Y_{\frac{n-1}{2}}] + 1) + \frac{1 - q^{(n+1)/2}}{1 - q^n}E[Y_{\frac{n+1}{2}}] && (6) \end{aligned}$$

Starting with $E[Y_1] = 0$, Equations (1), (4), and (6) give the recursion for $E[X_n]$.

Remark: The Halving Scheme is not Optimal

Although dividing a positive group into two subgroups of equal size (or as close to equal as possible when the group size is odd) seems reasonable, it may be surprising to know that it need not be optimal. For a counterexample, suppose that p is very small, so that except for n extremely large, a positive group of sized n will, with probability extremely close to 1, have exactly one positive member. Then, using

$$E[Y_n] = P(J=1)E[Y_n|J=1] + P(J=0)E[Y_n|J=0]$$

we obtain that

$$\begin{aligned}
E[Y_2] &\approx \frac{1}{2}2 + \frac{1}{2}1 = 3/2 \\
E[Y_3] &\approx \frac{1}{3}(1+1) + \frac{2}{3}(1+E[Y_2]) \approx 7/3 \\
E[Y_4] &\approx \frac{1}{2}(1+E[Y_2]+1) + \frac{1}{2}(1+E[Y_2]) \approx 3 \\
E[Y_5] &\approx \frac{2}{5}(1+E[Y_2]+1) + \frac{3}{5}(1+E[Y_3]) \approx 17/5
\end{aligned}$$

Thus, using our halving scheme

$$E[Y_8] \approx \frac{1}{2}(1+E[Y_4]+1) + \frac{1}{2}(1+E[Y_4]) \approx 9/2$$

On the other hand, if the positive group of size 8 were broken into subgroups of sizes 3 and 5, with the one of size 3 tested first, and the halving scheme followed from then on, then

$$E[\text{number}] \approx \frac{3}{8}(1 + E[Y_3] + 1) + \frac{5}{8}(1 + E[Y_5]) \approx 35/8$$

Thus, for small p, the split $(3, 5)$ is better than the split $(4, 4)$.

3 R code for $E[X_n]$

The following R code yields, when N is an even number and $q \in [0, 1]$, the value of a function $m(q, N)$ defined to equal the vector $(\text{argmin}_{n \leqslant N} \frac{E[X_n]}{n},\ \min_{n \leqslant N} \frac{E[X_n]}{n})$, where N is an even number and $q = 1 - p$. In the code $M[n] = E[Y_n]$, and $F[n] = \frac{E[X_n]}{n} = \frac{1+(1-q^n)E[Y_n]}{n}$.

```
> m = function (q, N){
+ M = array(0, N)
+ F = array(0, N)
+ M[1] = 0
+ M[2] = (2 + q)/(1 + q)
+ F[1] = 1
+ F[2] = (1 + (1 − q^2) * M[2])/2
+ c = N/2 − 1
+ for(k in 1 : c){
```

```
+   $M[2*k+1] = 1 + ((M[k]+1)*(1-q^k) + M[k+1]*(1-q^(k+1)))/(1-q^(2*k+1))$
+   $M[2*k+2] = (2 + q^(k+1) + 2*M[k+1])/(1+q^(k+1))$
+   $F[2*k+1] = (1 + (1-q^(2*k+1))*M[2*k+1])/(2*k+1)$
+   $F[2*k+2] = (1 + (1-q^(2*k+2))*M[2*k+2])/(2*k+2)\}$
+   $u =$ which.min$(F)$
+   $v = \min(F)$
+   $c(u, v)\}$
```

For given values of q, N, one inputs $m(q, N)$ and its value is given. For instance,

$m(.95, 1000)$	yields	13, 0.323212
$m(.98, 1000)$	yields	37, 0.166914
$m(.99, 1000)$	yields	75, 0.098020
$m(.999, 1000)$	yields	683, 0.014723
$m(.9999, 1000)$	yields	1000, 0.002479
$m(.9999, 5000)$	yields	4949, 0.001975
$m(.9999, 10, 000)$	yields	6827, 0.001971

When $p = .0001$, the preceding indicates that the optimal number to test when the upper limit is $5,000$ is 4949, and the optimal number is 6827 when the upper limit is $10,000$. This is quite interesting because, with

$F(n) = \frac{E[X_n]}{n}$, it shows that $F(6827) < F(4949) < F(5000)$, thus contradicting a seemingly reasonable hypothesis that $F(n)$ first decreases and then increases in n.

When p is unknown, we suggest that it be continually estimated by $\frac{x+1}{x+y+2}$ where x is the number of people known to be positive and y is the number known to be negative, and when a new group is to be tested its size be determined as if p were equal to its estimated value. This would mean that initially a single individual should be tested; if negative then the next group should be of size 2 and if positive of size 1, and so on.

4 Recursive Formula for Var(X_n)

With I and J as previously defined,

$$\text{Var}(X_n|I) \quad = \quad I\,\text{Var}(Y_n)$$

Using that $E[X_n|I] = 1 + I\,E[Y_n]$, the conditional variance formula yields

$$\begin{aligned}\text{Var}(X_n) &= E[\text{Var}(X_n|I)] + \text{Var}(E[X_n|I]) \\ &= (1-q^n)\text{Var}(Y_n) + E^2[Y_n]q^n(1-q^n) \qquad (7)\end{aligned}$$

Now, when $n \geqslant 2$ is even

$$\mathrm{Var}(Y_n|J) \quad = \quad \mathrm{Var}(Y_{\frac{n}{2}}) + J\,\mathrm{Var}(X_{\frac{n}{2}})$$

Thus, using (2) gives that for $n \geqslant 2$ even

$$\begin{aligned}\mathrm{Var}(Y_n) &= \mathrm{Var}(Y_{\frac{n}{2}}) + \frac{\mathrm{Var}(X_{\frac{n}{2}})}{1+q^{n/2}} + \frac{q^{n/2}\,E^2[X_{\frac{n}{2}}]}{(1+q^{n/2})^2} \\ &= \mathrm{Var}(Y_{\frac{n}{2}}) + \frac{(1-q^{n/2})\mathrm{Var}(Y_{\frac{n}{2}}) + E^2[Y_{\frac{n}{2}}]q^{n/2}(1-q^{n/2})}{1+q^{n/2}} + \frac{q^{n/2}\,E^2[X_{\frac{n}{2}}]}{(1+q^{n/2})^2} \\ &= \frac{2\,\mathrm{Var}(Y_{\frac{n}{2}}) + E^2[Y_{\frac{n}{2}}]q^{n/2}(1-q^{n/2})}{1+q^{n/2}} + \frac{q^{n/2}\,E^2[X_{\frac{n}{2}}]}{(1+q^{n/2})^2} \qquad (8)\end{aligned}$$

When $n \geqslant 3$ is odd

$$\mathrm{Var}(Y_n|J) \quad = \quad J\left(\mathrm{Var}(Y_{\frac{n-1}{2}}) + \mathrm{Var}(X_{\frac{n+1}{2}})\right) + (1-J)\mathrm{Var}(Y_{\frac{n+1}{2}})$$

Hence, using (5), we obtain that when $n \geqslant 3$ is odd

$$\begin{aligned}\mathrm{Var}(Y_n) &= \frac{1-q^{(n-1)/2}}{1-q^n}\left(\mathrm{Var}(Y_{\frac{n-1}{2}}) + \mathrm{Var}(X_{\frac{n+1}{2}}) - \mathrm{Var}(Y_{\frac{n+1}{2}})\right) + \mathrm{Var}(Y_{\frac{n+1}{2}}) \\ &\quad + \left(E[Y_{\frac{n-1}{2}}] + E[X_{\frac{n+1}{2}}] - E[Y_{\frac{n+1}{2}}]\right)^2 \frac{(1-q^{(n-1)/2})(q^{(n-1)/2}-q^n)}{(1-q^n)^2} \\ &= \frac{1-q^{(n-1)/2}}{1-q^n}\left(\mathrm{Var}(Y_{\frac{n-1}{2}}) - q^{(n+1)/2}\mathrm{Var}(Y_{\frac{n+1}{2}}) + E^2[Y_{\frac{n+1}{2}}]q^{(n+1)/2}(1-q^{(n+1)/2})\right) \\ &\quad + \mathrm{Var}(Y_{\frac{n+1}{2}}) + \left(E[Y_{\frac{n-1}{2}}] + E[X_{\frac{n+1}{2}}] - E[Y_{\frac{n+1}{2}}]\right)^2 \frac{(1-q^{(n-1)/2})(q^{(n-1)/2}-q^n)}{(1-q^n)^2} \\ &\qquad (9)\end{aligned}$$

Starting with $\mathrm{Var}(Y_1) = 0$, Equations (7), (8), and (9), along with the previous recursion for determining $E[Y_n]$, give the recursion for computing $\mathrm{Var}(X_n)$.

References

[1] Eugene Litvak, Xin M. Tu and Marcello Pagano, “Screening for the Presence of a Disease by Pooling Sera Samples," *Journal of the American Statistical Association*, Vol. 89, No. 426, pp. 424-434, 1994

[2] Haran Shani-Narkiss, Omri David Gilday, Nadav Yayon, Itamar Daniel Landau, “Efficient and Practical Sample Pooling for High-Throughput PCR Diagnosis of COVID-19," preprint , 2020

[3] Ron Brookmeyer, “Analysis of Multistage Pooling Studies of Biological Specimens for Estimating Disease Incidence and Prevalence." *Biometrics,* Vol. 55, No. 2, pp. 608-612, 1999