Dorfman-Rosenblatt Testing

Nick Landolfi and Sanjay Lall
Stanford University

Revision: 192a892

Motivation

- in 2021: asymptomatic mass testing for COVID
  - have a large population of people
  - want to determine the individuals with COVID
  - polymerase chain reaction (PCR) tests are expensive and take time
- in the 1940s: US Public Health Service and Selective Service screens recruits for syphilis
  - have a sensitive antigen-based blood test
  - want to determine individuals with syphilis
  - reference: *The Detection of Defective Members of Large Populations*, Dorfman 1943
    - anecdotally, Robert Dorfman discussed with David Rosenblatt

Basic idea

- the usual procedure: test all n people
  - requires n tests
  - call this *individual testing* or *non-grouped testing* (in contrast with below)
- a better idea: split each blood sample in half, pool blood from 5 individuals into single test
  - if a pool is negative, then declare all individuals in the pool negative
  - if a pool is positive, then separately re-test each individual in the pool (using other $1/2$ of blood)
  - call this procedure *group testing* or *pooled testing*
  - when does it help? how does choice of group size matter?

Model

- have a background probability space $(\Omega, \mathcal{A}, \mathbf{P})$
  - with n i.i.d. Bernoulli random variables $x_i : \Omega \to \{0, 1\}$
  - call $p = \mathbf{P}[x_i = 1]$ the *prevalence rate*
- we want, for an $\omega \in \Omega$, to determine the non-zero elements of

$$x(\omega) = (x_1(\omega), \ldots, x_n(\omega)) \in \{0, 1\}^n$$

- we want to minimize the expected number of tests required
  - i.e., we care about average case over $\mathbf{P}$
  - called the *probabilistic* setting

Analysis

- if we have n individuals and a *pool size* of m
- then we have $\lceil n/m \rceil$ *pooled tests*, since n/m need not be an integer
  - gives $\lfloor n/m \rfloor$ *full pools* of size m and (possibly) 1 *partial pool* of size $\mathrm{mod}(n,m)$
- for a pool of size k, all individuals are negative with probability $(1-p)^k$
  - so the pool is positive with probability $1-(1-p)^k$
  - we retest all individuals in a pool if it is positive
- the expected total number of tests is (if, in case $\mathrm{mod}(n,m)=1$, we still retest)

$$T(n,m,p) = \underbrace{\lceil n/m \rceil}_{\#\text{ pools}} + \underbrace{\lfloor n/m \rfloor m(1-(1-p)^m)}_{\text{full pools retesting}} + \underbrace{\mathrm{mod}(n,m)(1-(1-p)^{\mathrm{mod}(n,m)})}_{\text{partial pool retesting}}$$

  - compare with no pooling: expected number of tests is (constant) n
  - obtain *cost per individual* by dividing by n

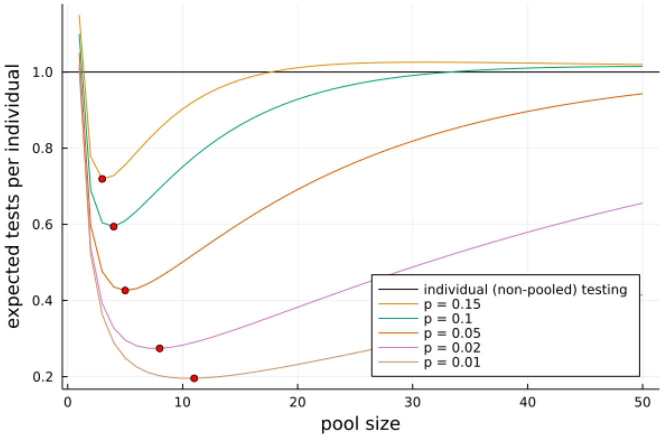
Idealization as $n \to \infty$

- define *idealized expected tests per individual* by dividing by n and taking $n \to \infty$
- we have that $\lim_{n\to\infty} {}^1\!/_n T(n, m, p)$ is ${}^1\!/_m + 1 - (1-p)^m$
  - denote by $T_\infty(m, p)$
  - use $\lim_{n\to\infty} {}^1\!/_n \lceil {}^n\!/_m \rceil = \lim_{n\to\infty} {}^1\!/_n \lfloor {}^n\!/_m \rfloor = {}^1\!/_m$ and $\lim_{n\to\infty} {}^1\!/_n \operatorname{mod}(n, m) = 0$
  - theory says: fix p, and then optimize m to minimize this expression
  - removes dependence on n, keeps dependence on p
- two interpretations:
  - can interpret as the *average number of tests per individual* as n tends large
  - can interpret as the *relative cost compared with non-grouped testing* as n tends large

Computing optimal pool size, given prevalence

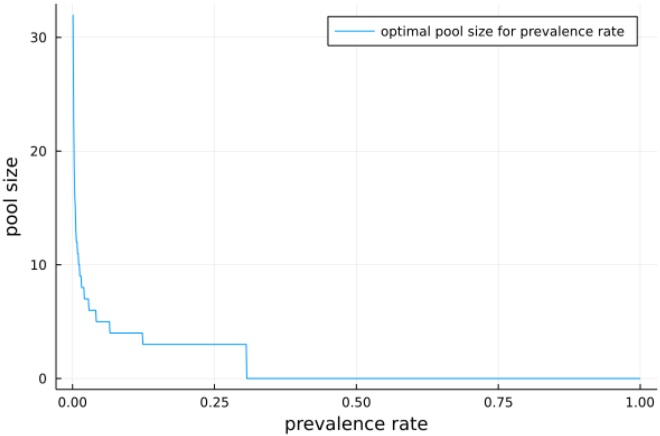
- given p and n, can find $m \in \mathbf{Z}$, $0 < m < n$ to minimize cost per individual $T(n, m, p)$
- or, can fix p and find $m \in \mathbf{Z}$, $m > 0$ to minimize idealized cost for individual $T_\infty(m, p)$
- or, can relax to $m \in \mathbf{R}$, $m > 0$ and optimize one-dimensional function ${}^1/_m + 1 - (1-p)^m$

Expected tests vs. pool size for various p



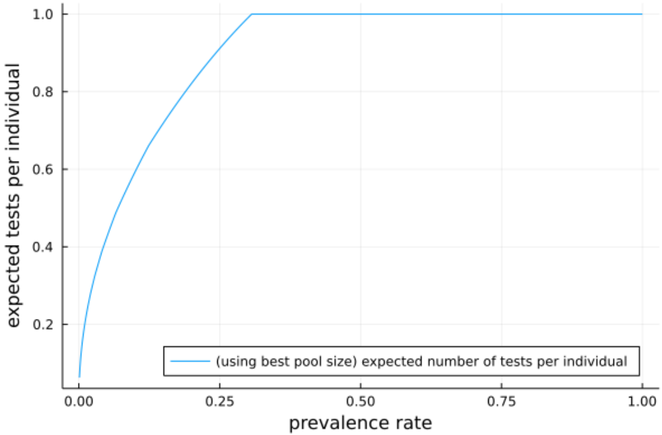
- want to beat 1 test per individual; red dots indicate optima at different prevalence levels
- if we have a small prevalence, larger pools lead to more savings (up to a point)

Optimal pool size vs. prevalence



- as prevalence p increases, best pool size decreases; small p indicates large pools
- however, prevalence can be so large that pooling is suboptimal

Indicated expected cost per individual vs. prevalence



- suppose you use best pool size for each prevalence rate, here is the expected number of tests

Parallel pooling

- *double-pooling* idea: pick pool size m and split into random pools twice
  - *re-test* an individual if and only *if both* of their pools are *positive*; the expected cost

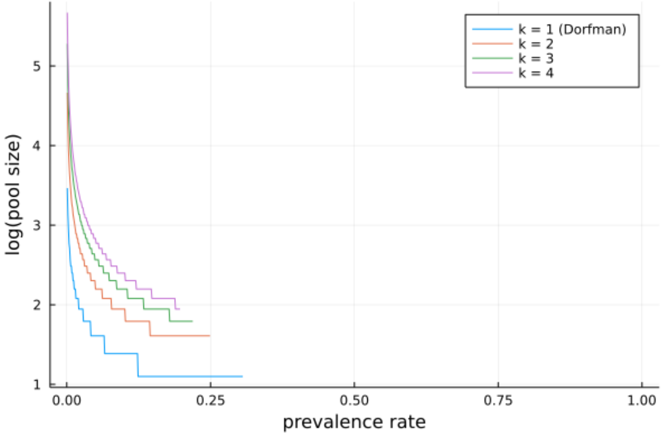
$$\underbrace{2\lceil n/m\rceil}_{\#\text{ pools}} + \mathbf{E}(\text{number of individuals in two positive groups})$$

- expectation is n times $\mathbf{E}$(individual is in 2 positive groups), denote $1-p$ by q and deduce
  - in both full groups: $\left(\frac{\lfloor n/m\rfloor m}{n}\right)^2\left(p+q(1-q^{m-1})^2\right) \to p+q(1-q^{m-1})^2$ as $n\to\infty$
  - in one full and one partial: $2\frac{\mathrm{mod}(n,m)m\lfloor n/m\rfloor}{n^2}\left(p+q(1-q^{m-1})(1-p)^{\mathrm{mod}(n,m)-1}\right) \to 0$ as $n\to\infty$.
  - in both partial groups: $\left(\frac{\mathrm{mod}(n,m)}{n}\right)^2\left(p+1-q^{\mathrm{mod}(n,m)-1}\right)^2 \to 0$ as $n\to\infty$
- can use these formulae to find expected cost; or can idealize as $n\to\infty$ (as before) and obtain

$$2/m + p + (1-p)(1-(1-p)^{m-1})^2$$

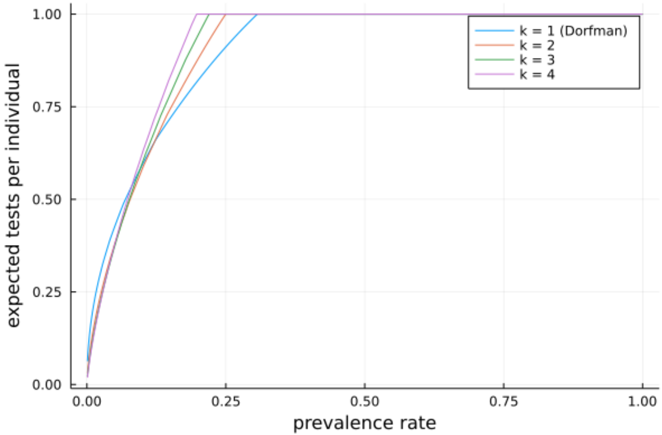
- this method is proposed in *A Note on Double Pooling Tests*, Broder & Kumar 2020
  - generalizes to *k-parallel pooling*: one obtains idealize relative cost $k/m + p + (1-p)(1-(1-p)^{m-1})^k$

Parallel testing: log optimal pool size vs. prevalence



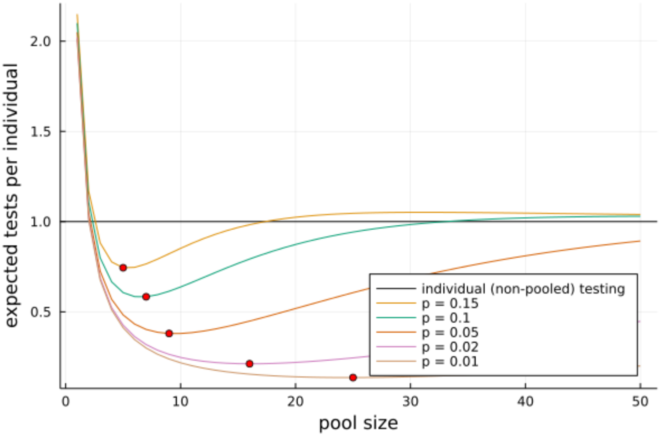
- pools get larger as parallelism increases
- prevalence rate at which you stop pooling decreases with parallelism

Parallel testing: indicated expected cost vs prevalence



- notice that double pooling (and others) beat single (Dorfman) pooling at low prevalence

Double pooling: expected tests vs. pool size for various p



- compare with earlier plot for Dorfman pooling; pool sizes are larger here