Efficient Disease Screening Using Group Testing and Symmetric Probability

Nick Landolfi
Stanford University

Revision: 24 October 2022

Research so far...

- robot reward learning from demonstrations and preferences
- multi-task model-based reinforcement learning
- data center anomaly detection and sparse structural equation model learning
- **this talk:** group testing for symmetric distributions

Outline

1. Group testing
2. Symmetric distributions
3. Algorithm
4. Real world example
5. Future work

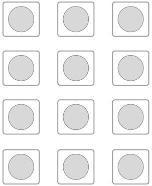
Group testing to save resources

Group testing to save resources

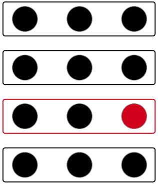
- we have a batch of n specimens to screen for a binary trait
  - have blood draws, want to screen for syphilis using antigen tests
  - have nasal swabs, want to screen for COVID using RT-PCR tests
  - have liquid biopsies, want to screen for cancer using ct-DNA tests
- we want to know trait associated with each specimen
- **basic idea**: pool specimens together in groups of size $k > 1$, test as a group

Saving tests by choosing groupings

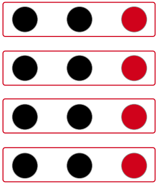
population of size $n = 12$



12 individual tests
(does not depend on outcomes)



4 group tests, 3 retests
7 tests used
group tests help



4 group tests, 12 retests
16 tests used
group tests hurt

if we knew the distribution, we could design groupings that minimize expected cost

Group testing and Dorfman's procedure

- we may test several specimens together as a *group*, and observe that either
  1. *all* the specimens are negative *or*
  2. *at least* one of the specimens is positive
- Dorfman[1] proposed an adaptive two-stage procedure
  - pool specimens into groups of size $k > 1$, each group is tested
    - if the group tests negative, declare all k specimens negative, saving $k - 1$ tests
    - if the group tests positive, retest all specimens in the group individually
  - **punchline**: if most groups tests negative, pooling saves tests
  - benefits: simple, parallel, only split sample into two portions

[1] *The detection of defective members of large populations*, Annals of Mathematical Statistics, 1943

Minimizing expected number of tests

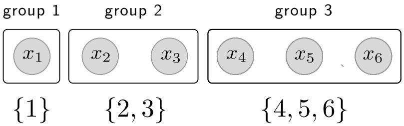
- n individuals
- $x = (x_1, \ldots, x_n)$ where binary random variable x_i is the *status* of individual i
- partition $\{1, \ldots, n\}$ into *grouping* $G = \{H_1, \ldots, H_k\}$ where $H_i \subset \{1, \ldots, n\}$ is the ith *group*
- expected number of tests is $\mathbf{E}C(G, x) = \sum_{H \in G} \mathbf{E}T_H(x)$ where

$$T_H(x) = \begin{cases} 1 & \text{if } x_i = 0 \text{ for all } i \in H \\ 1 + |H| & \text{if } x_i = 1 \text{ for some } i \in H \end{cases}$$

  - $T_H(x)$ is number of tests used for group H

Minimizing expected number of tests: example

- for example $n = 6$, and we partition into three groups



- expected number of tests is

$$\underbrace{1}_{\text{group 1}} + \underbrace{1 + 2\,\text{prob}(x_2 = 1 \text{ or } x_3 = 1)}_{\text{group 2}} + \underbrace{1 + 3\,\text{prob}(x_4 = 1 \text{ or } x_5 = 1 \text{ or } x_6 = 1)}_{\text{group 3}}$$

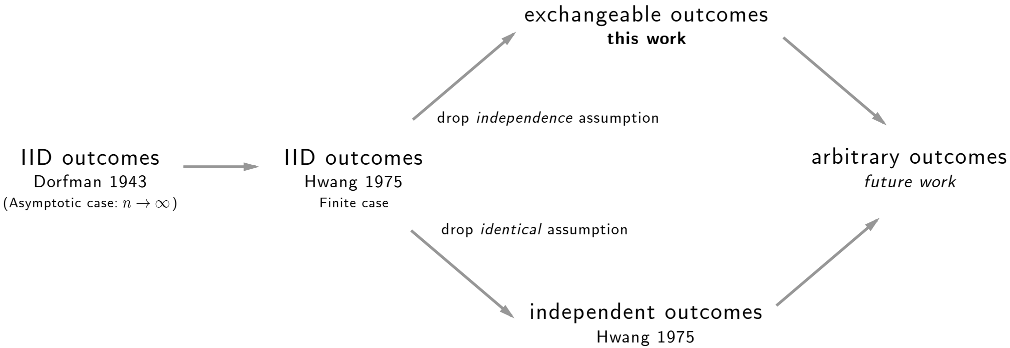
  - always need 3 group tests, may need additional individual tests

Minimizing expected number of tests: problem

- $x_1, \ldots, x_n$ have distribution $p : \{0,1\}^{\{1,\ldots,n\}} \to [0,1]$
- **Problem.** given p, find a partition G of $\{1, \ldots, n\}$ to minimize the expected number of tests
  - efficient algorithms when $x_1, \ldots, x_n$ are IID or just independent[2]
  - our work: efficient algorithm when $x_1, \ldots, x_n$ are *exchangeable*
    - roughly means any subset has the same distribution
    - allows modeling correlation in test outcomes

[2] Hwang, *A generalized binomial group testing problem*, Journal of the American Statistical Association, 1975

Problem 1: overview of assumptions on $x_1, \ldots, x_n$



Symmetric distributions

Rearranging distributions and definition of symmetry

- given outcomes $x \in \{0,1\}^{\{1,\dots,n\}}$ and permutation g of $\{1,\dots,n\}$
- rearrange x as usual via composition $x \circ g$
- likewise, rearrange distribution p to distribution $p^g : \{0,1\}^{\{1,\dots,n\}} \to [0,1]$ defined by

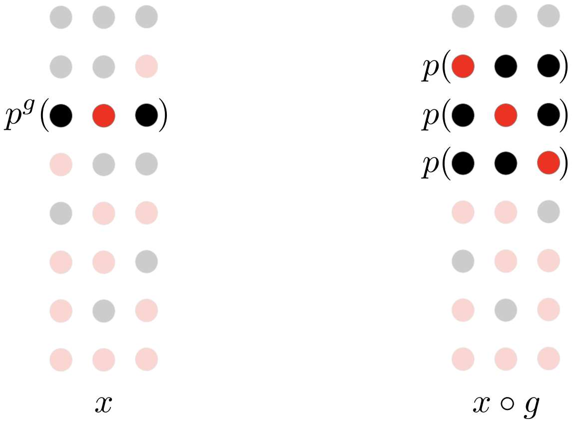
$$p^g(x) = p(x \circ g)$$

- call p *symmetric* if

$$p = p^g \quad \text{for all permutations } g \text{ of } \{1,\dots,n\}$$

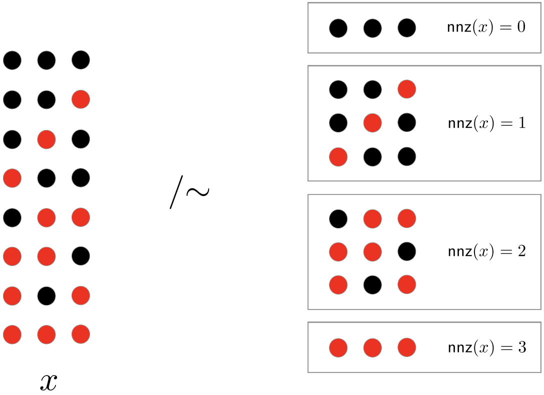
    - alternative language: call $x_1,\dots,x_n$ *exchangeable*
- p is a *permutation-invariant* function

Rearranging distributions



consider g swapping 1 and 3; symmetry means that all these probabilities are the same

Symmetric distributions are constant on equivalence classes



permutations give equivalence relation; $\mathsf{nnz}(x)$ is number of nonzero values of x

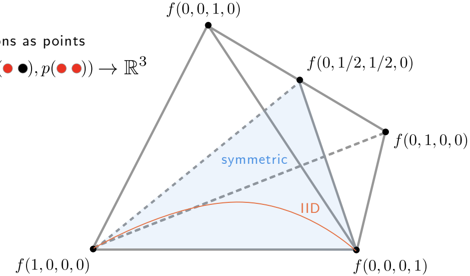
Examples of symmetric distributions

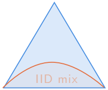
- any IID distribution is symmetric
- any mixture (convex combination) of symmetric distributions is symmetric
- simple random sampling produces symmetry
- shuffling creates symmetry

Geometry of symmetric distributions

embed distributions as points

$f : (p(\bullet\,\bullet), p(\bullet\,{\color{red}\bullet}), p({\color{red}\bullet}\,\bullet), p({\color{red}\bullet}\,{\color{red}\bullet})) \to \mathbb{R}^3$





set corresponding to all distributions is tetrahedron, that to all symmetric distributions is 2D simplex

Symmetric marginals

- **Fact:** Suppose $p : \{0,1\}^{\{1,\ldots,n\}} \to [0,1]$ is a distribution. Then

$$p \text{ is symmetric} \iff p_H = (p_J)^g \text{ for all bijections } g : J \to H \text{ where } H, J \subset P$$

  - p_H is the *marginal* over the variables $\{x_i\}_{i \in H}$
- has two intuitive interpretations
  - says that *all marginals of a symmetric distribution are symmetric*
    - i.e., any subset of exchangeable random variables is exchangeable
  - says that *all same-size marginals of a symmetric distribution agree*
    - e.g., the distribution of any three test outcomes is the same

Representation via marginals

- **Fact:** Suppose $p : \{0,1\}^{\{1,\ldots,n\}} \to [0,1]$ is a distribution. Then p is symmetric if and only if there exists a function $q : \{0,1,\ldots,n\} \to [0,1]$ such that

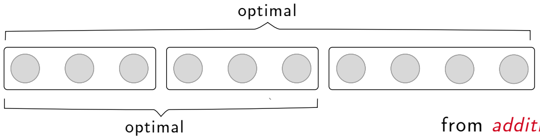
$$p_H(\mathbf{0}) = q(|H|) \quad \text{for all } H \subset \{1,\ldots,n\}$$

- q is a nonobvious *representation* for a symmetric distribution
  - $q(h)$ is the probability that a group of size h tests negative
  - q is the input representation to our algorithm

Main result and algorithm

Optimal partitions have optimal substructure

- motivation for a dynamic programming approach
- **Fact:** any subset of an optimal partition is optimal *for the subpopulation it partitions*

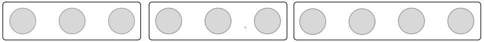


from *additive* cost

Simplifications under symmetry

- for *symmetric* distributions...
- (1) the cost of a group depends only on its size, denote by T_h for group of size h
- thus, (2) the cost of a grouping only depends on the number of groups it has of each size
  - depends on *pattern* π of a grouping where $\pi(h)$ is the number of groups of size h

cost is $T_3 \quad + \quad T_3 \quad + \quad T_4$



pattern is $\pi(1) = 0, \;\; \pi(2) = 0, \;\; \pi(3) = 2, \;\; \pi(4) = 1, \;\; \pi(5) = 0, \;\; \ldots$

- hence, (3) size-m subpopulations have same optimal patterns, same optimal cost $C^\star_m$

Algorithm and main result

- Fact: If $x_1, \ldots, x_n$ have *symmetric* distribution p, then

$$C_m^\star = \min_{h=1,\ldots,m} \{C_{m-h}^\star + T_h\} \quad \text{for all } m = 1, \ldots, n$$

  - where $C_m^\star$ optimal cost of subpopulation of size m and T_h is cost of testing group of size h

- Algorithm: to compute $C_1^\star, \ldots, C_n^\star$ and optimal patterns $\pi^1, \ldots, \pi^n$
  - take π^1 so that $\pi_1^1 = 1$ and $\pi_n^1 = 0$ for $n \neq 1$, take $C_1^\star = T_1$
  - for $k = 2, \ldots, n$, find h_k a minimizer of $f(h) = C_{k-h}^\star + T_h$, define π_k by

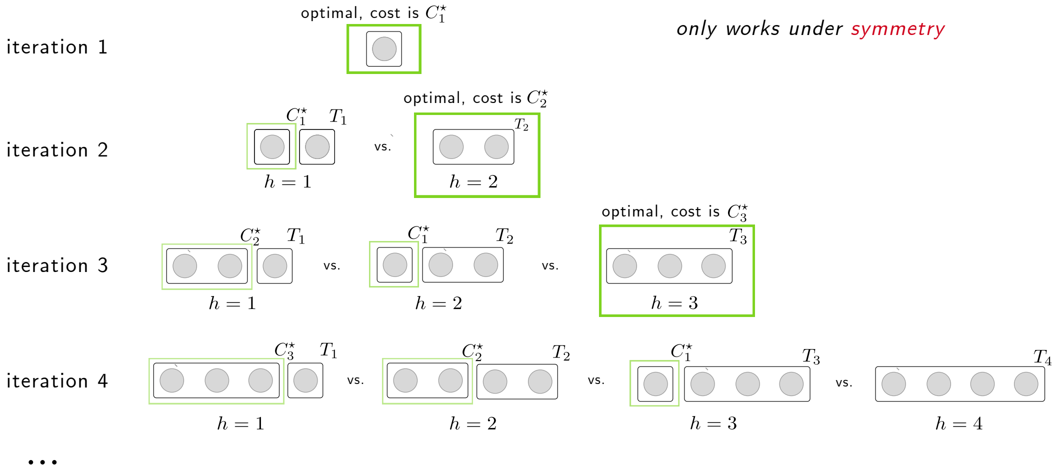
$$\pi_k(j) = \begin{cases} \pi_{k-h_k}(j) + 1 & \text{if } j = h_k \\ \pi_{k-h_k}(j) & \text{otherwise} \end{cases}$$

    and take $C_k^\star = C_{k-h_k}^\star + T_{h_k}$

- **Theorem:** partitions computed in this way are optimal

Algorithm visualization

only works under *symmetry*



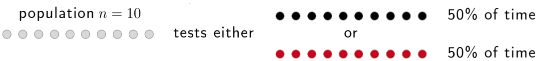
Simulation and data fitting

Comparisons

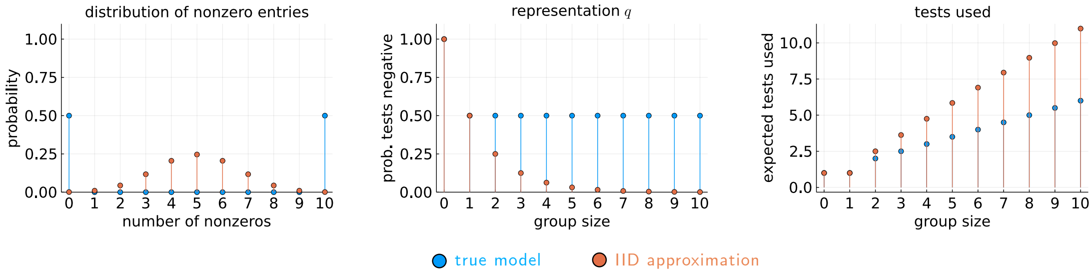
- for simulation and a real dataset, compare different approaches
  - prior tools, assuming IID outcomes; infinite (Dorfman) and finite (Hwang) cases
  - tool we built, assuming exchangeability
- in some cases, different approaches indicate the same pooling
- for intuition, we show examples where the indicated poolings are different

Example 1: 10 individuals, all or none positive

- simple extreme example for intuition



  - at prevalence of 1/2, both IID-∞ and IID-finite say test individually (10 tests)
  - symmetric says pool one group of size 10 (6 tests on avg.)



Approximation by symmetric distributions and fitting

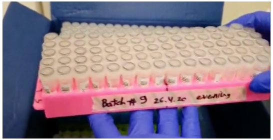
- Problem: given arbitrary distribution $r : \{0,1\}^{\{1,\dots,n\}} \to [0,1]$, find a distribution p to

$$\begin{array}{rl} \text{minimize} & d_{kl}(r,p) \\ \text{subject to} & p \text{ is symmetric} \end{array}$$

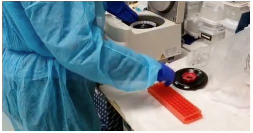
- Solution: pick the symmetric distribution which puts the same mass on equivalence classes as r
  - indicates solution to *maximum likelihood estimation*
  - count number of samples with no positives, one positive, two positives, and so on...

Barak et al. dataset 2021 methodology and observation

1. batch of 80 arrives

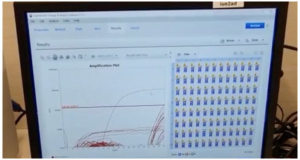


2. spin down lysate

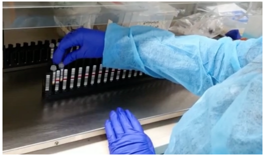


3. robot pools/mixes samples

4. RT-PCR test (up to 90 pools in parallel)



5. individual retesting



- “*in reality, samples arrive in batches: from colleges, nursing homes, or health care personnel...thereby increasing the number of positive samples*”[3]

[3] Barak et al., *Lessons from applied large-scale pooling of 133,816 SARS-CoV-2 RT-PCR tests*, 2021

Barak et al. 2020: our results

- take first 2 months of data (prevalence stable, about 0.2%)
  - corresponds to 500 batches of size 80; fit on first half, test on second half
- group testing should help at low prevalence
  - individual testing uses **40,000 tests**
  - Barak et al. partition 8, 8, 8, 8, 8, 8, 8, 8, 8, 8; uses 2940 tests
  - IID model indicates partition 20, 20, 20, 20; uses 1,660 tests
  - symmetric model indicates partition 27, 27, 26; uses **1,630 tests**

Additional topics not discussed and future work...

- characterize formally when symmetry helps
- use sampling to reduce number of tests (as in example 1)
- use features to learn the probability a sample will test positive
- use permutation invariant models to learn probability a group with some set of features will test positive

Efficient disease screening using group testing and symmetric probability

- we generalized classical group testing to symmetric distributions
- demonstrated a proof of concept on real data

Thank you!

Extra slides

An infectious disease example: group exposure model

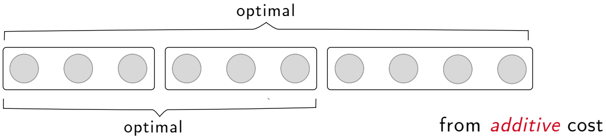
- set of symmetric distributions is convex
- given symmetric distributions r and s along with a mixing parameter μ in $[0, 1]$, define

$$p(x) = (1 - \mu)r(x) + \mu s(x)$$

  - interpret p as modeling outcomes that depend on some *unobserved* event
    - latent event occurs with probability μ
  - call $\mathbf{E} \sum_{i=1}^{n} x_i/n$ the *prevalence rate*
    - if r and s have prevalence rates ρ_r and ρ_s, then p has rate $(1 - \mu)\rho_r + \mu\rho_s$
    - if $\rho_s > \rho_r$ we may say the unobserved *exposure* event *increases* the prevalence
  - straightforward generalization to ℓ levels, Bayesian interpretation of mixing parameters

Optimal partitions have optimal substructure

- motivation for a dynamic programming approach
- call a partition $F^\star$ of $S \subset P$ *optimal* if $\mathbf{E}C(F^\star, x) \leq \mathbf{E}C(F, x)$ for all other partitions F
- **Fact:** If $F^\star$ is optimal for S, then for any $E \subset F^\star$, E is optimal for $\cup E$
  - any subset of an optimal partition is optimal *for the subpopulation it partitions*



Tests used for a group depends only on size

- for any distribution we have

$$\mathbf{E}T_H(x) = \begin{cases} 1 & \text{if } |H| = 1 \\ 1 + |H| \operatorname{Prob}(S_H(x) = 1) & \text{otherwise} \end{cases}$$

- if p is *symmetric*, we can express the second case

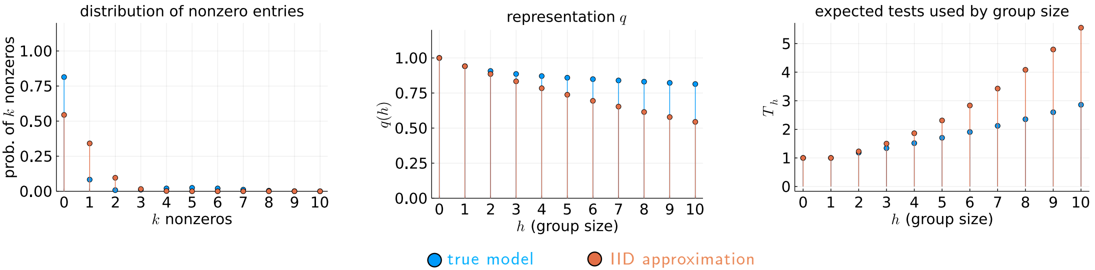
$$\begin{aligned} 1 + |H| \operatorname{Prob}(S_H(x) = 1) &= 1 + |H|(1 - \operatorname{Prob}(S_H(x) = 0)) \\ &= 1 + |H|(1 - p_H(\mathbf{0})) \\ &= 1 + |H|(1 - q(|H|)) \end{aligned}$$

  - the right hand side depends only on $|H|$

- not true without symmetry: for example, independent outcomes with different probabilities

Example 2: group exposure

- simple for intuition: w.p. 0.9, prevalence 0.01, w.p. 0.1 prevalence 0.5
  - the population prevalence is 0.059
  - IID-∞, IID-finite: two pools of 5 (3.41 tests on avg.), symmetric: one pool of size 10 (2.85 tests on avg.)



Example 3: multi group exposure

- here we have $n = 30$, we concatenate three of the group exposure models each of size 10
    - exposure model same as before, 90% of time IID with prevalence 0.01, 10% of time IID with prevalence 0.5
- draw 10^5 samples, and fit a distribution using methodology on previous slide
- IID, finite and infinite, indicates partition 5, 5, 5, 5, 5, 5; uses 10.2 tests on average
- symmetric indicates partition 8, 8, 7, 7; uses 9.8 tests on average