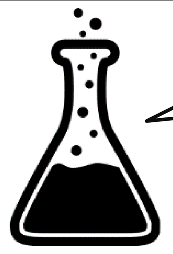
Estimating the Number and Effect Sizes of Non-null Hypotheses

Jennifer Brennan, Ramya Korlakai Vinayak, Kevin Jamieson

{jrb, ramya, jamieson}@cs.washington.edu

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Begin with a scientific question



Which of 10,000 *Drosophila* genes inhibit virus growth?

☐ **Option 1** Full Experiment

Find all genes that inhibit virus replication by at least 2x (the *effect size*), as measured by a fluorescent reporter

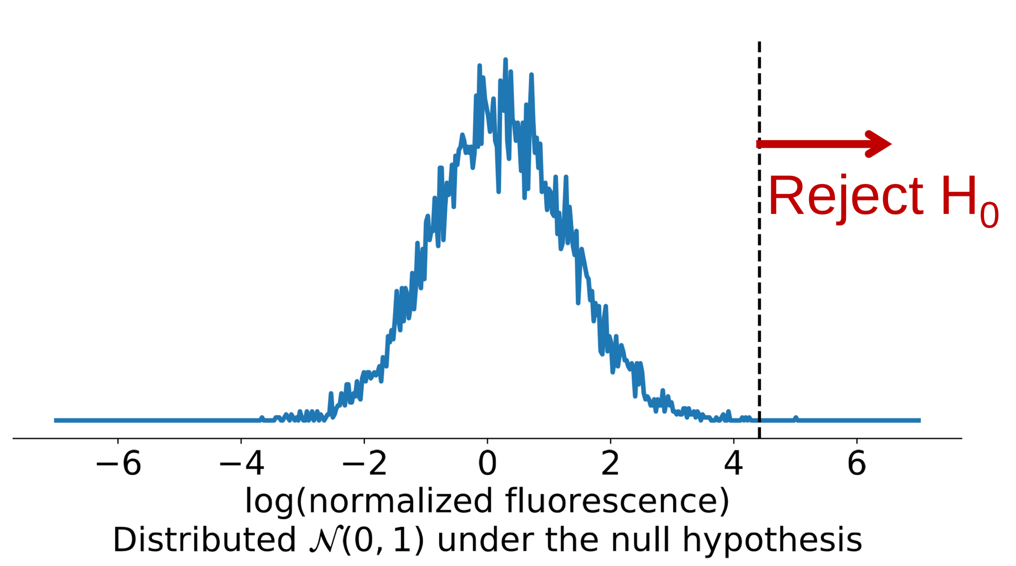
Replicates: *Cost: $$$$*

☑ **Option 2** Pre-screen

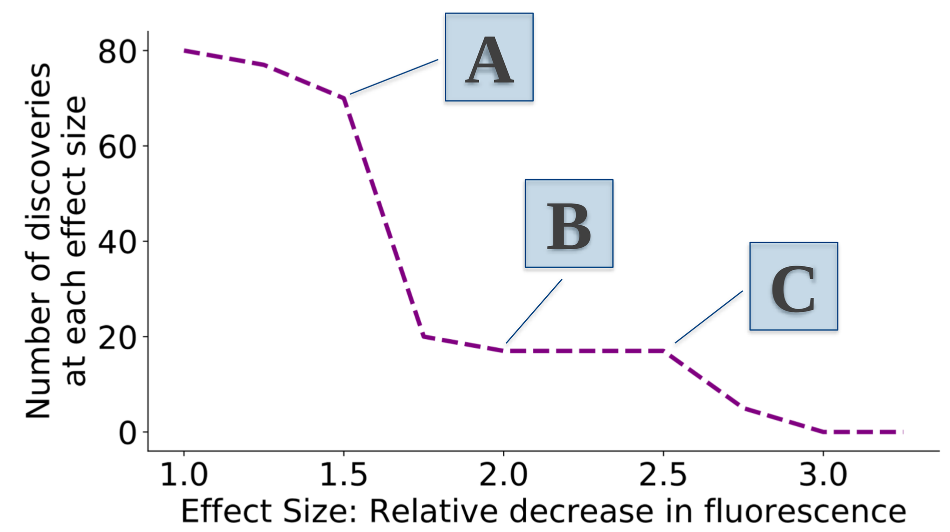
Count the number of genes with each effect size

Replicates: *Cost: $*

Prescreen with few replicates per gene



Multiple hypothesis testing identifies no discoveries



Our estimator indicates discoveries exist

Choose an experimental design

A Find all genes with effect size at least 1.5x

Guaranteed yield: At least 70 genes

Replicates: *Cost: $$$$$*

B Find all genes with effect size at least 2x

Guaranteed yield: At least 17 genes

Replicates: *Cost: $$$$*

C Find all genes with effect size at least 2.5x

Guaranteed yield: At least 17 genes

Replicates: *Cost: $$$*

Question When testing multiple hypotheses simultaneously, can we <u>estimate the number</u> of positive effect sizes (discoveries) with fewer samples than it takes to <u>identify</u> those discoveries?

Our Answer Yes. If we are testing *n* hypotheses and the effect sizes are small, then we can detect the existence of discoveries with <u>*n* times fewer samples</u> than we would need for identification.

Problem Statement

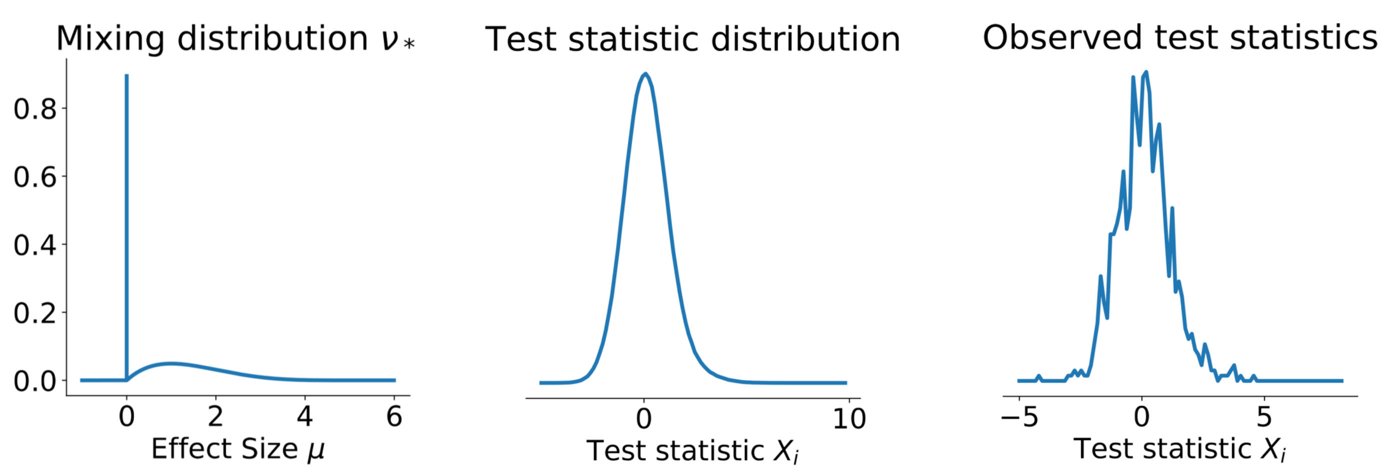
Formal Statement

Let ν_* be a distribution on $\mathbb{R}$. Let f_μ be the known parametric distribution of test statistics with effect size μ, e.g. $f_\mu = \mathcal{N}(\mu, 1)$. Draw $\{\mu_i\}_{i=1}^n \sim \nu_*$ (unobserved) and $\{X_i\}_{i=1}^n \sim f_{\mu_i}$ (observed).

Goal

Estimate, for all $\gamma \in \mathbb{R}$ and without overestimating,

$$\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$$



Our Estimator

Let

$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le t\}$ be the empirical CDF

$F_\nu(t) = \mathbb{P}_{\mu\sim\nu,\, X\sim f_\mu}(X \le t)$ be the true CDF under ν

For any $\gamma \in \mathbb{R}$, for a specified probability of overestimation α, our estimator is given by

$$\widehat{\zeta}_n(\gamma) = \min_{\nu: \|\widehat{F}_n - F_\nu\|_\infty \le \sqrt{\frac{\log(2/\alpha)}{2n}}} \int_\gamma^\infty \nu(x)dx. \qquad (1)$$

Theoretical Results

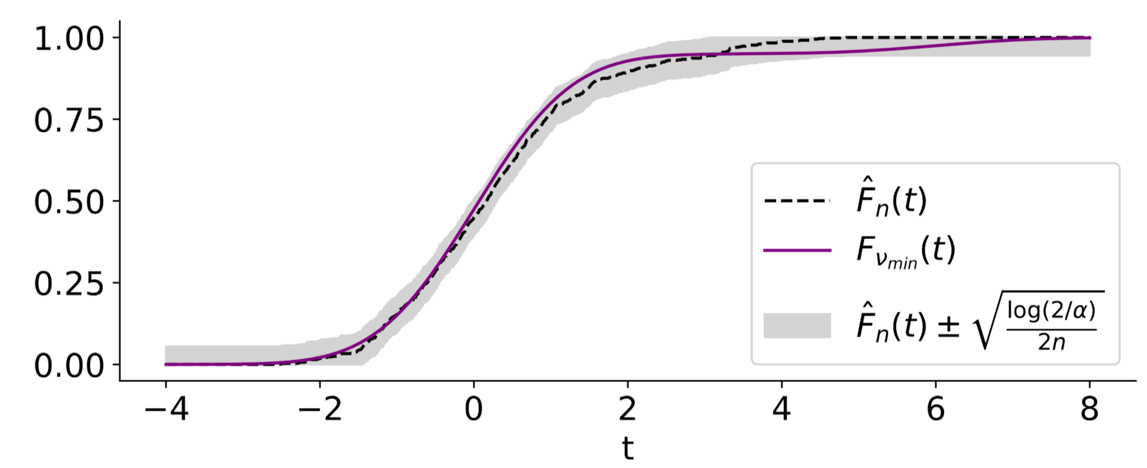
Theorem. *For $i = 1, 2, \ldots, n$, let $\mu_i \sim \nu_*$ and $X_i \sim f_{\mu_i}$ where each draw is iid. Let our simultaneous estimator be given by (1). Then,*

$$\mathbb{P}\left(\exists\gamma : \widehat{\zeta}_n(\gamma) > \zeta_{\nu_*}(\gamma)\right) \le \alpha.$$

Furthermore, with probability at least $1-\delta$, for all $\gamma \in \mathbb{R}$ and $\varepsilon \in (0, \zeta_{\nu_*}(\gamma)]$ we have $\zeta_{\nu_*}(\gamma) - \widehat{\zeta}_n(\gamma) \le \varepsilon$ whenever

$$n \ge \frac{\log\left(\frac{4}{\alpha\delta}\right)}{\left(\min_{\nu:\mathbb{P}_\nu((\gamma,\infty))\le\zeta_{\nu_*}(\gamma)-\varepsilon} \|F_\nu - F_{\nu_*}\|_\infty\right)^2}.$$

Intuition and Proof Sketch

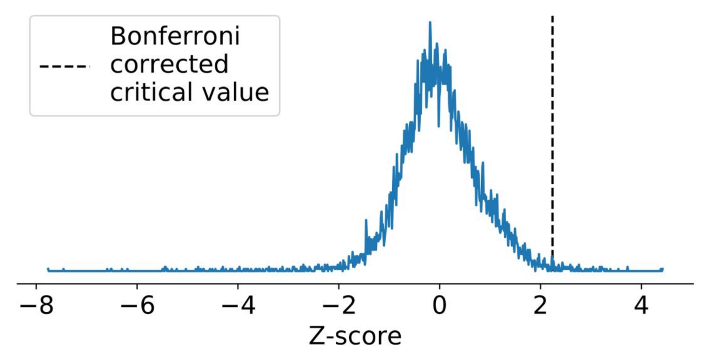


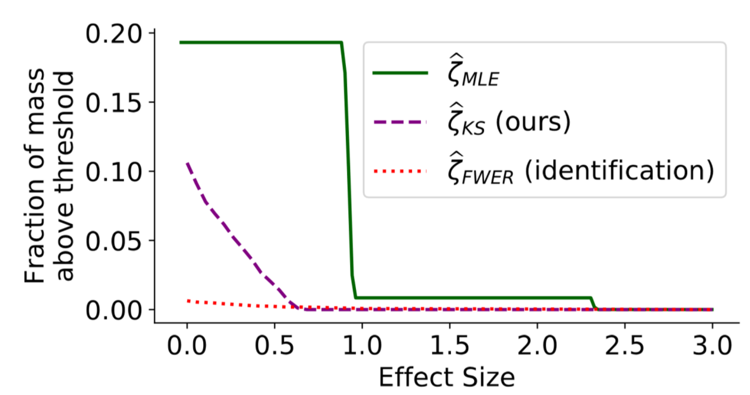
By the **DKW inequality**, the true CDF F_{ν_*} is contained in the ℓ_∞ ball around $\widehat{F}_n$. If we find the ν that stays inside this ball but has the *least* mass above γ (the minimizer of (1), shown by $F_{\nu_{min}}$), then with high probability, this is a lower bound on $\zeta_{\nu_*}(\gamma)$.

The sample complexity result follows from the DKW inequality and the triangle inequality. We want to bound $\|F_{\nu_*} - \widehat{F}_n\|_\infty$ subject to the constraints on ν in (1), so we control it using the DKW bound on $\|F_{\nu_*} - \widehat{F}_n\|_\infty$, and the constrained quantity $\|F_\nu - F_{\nu_*}\|_\infty$ found in the theorem statement.

Experimental Results

Real Data





Data Two z-scores from gene knock-out experiments on each of 13,071 *Drosophila* genes [1]

Results

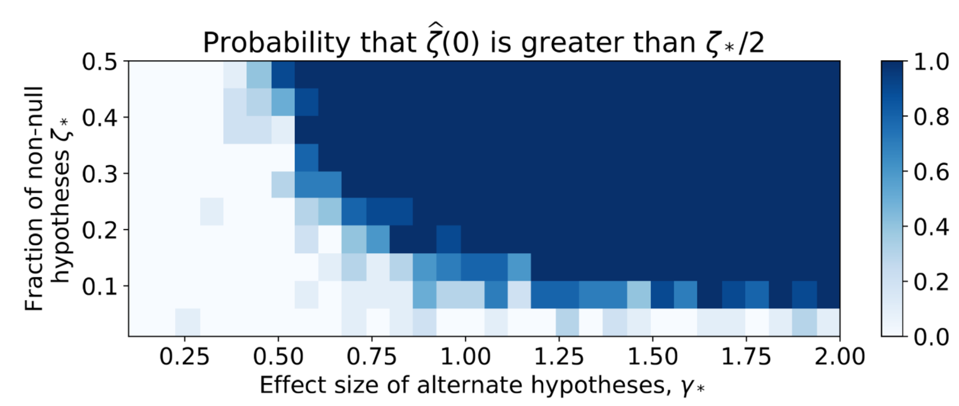
- Our estimator finds evidence of more discoveries than multiple testing can identify, especially at small effect sizes
- MLE suggests there may be even more discoveries – but the MLE frequently overestimates

[1] Linhui Hao et al. "Drosophila RNAi screen identifies host genes important for influenza virus replication". In: *Nature* 454.7206 (2008), p. 890.

Synthetic Data

Gaussian test statistics

Let $\nu_* = (1-\zeta_*)\delta_0 + \zeta_*\delta_{\gamma_*}$ and $f_\mu = \mathcal{N}(\mu, 1)$, as when the data are z-scores. The following figure shows the probability of detecting at least half of the discoveries, as a function of both γ_* and ζ_*, for a fixed $n = 10^4$.



Poisson test statistics

Our estimator also works on non-Gaussian data, as demonstrated here.