Estimating the number and effect sizes of non-null hypotheses

Jennifer Brennan, Ramya Korlakai Vinayak, Kevin Jamieson

jrb@cs.washington.edu

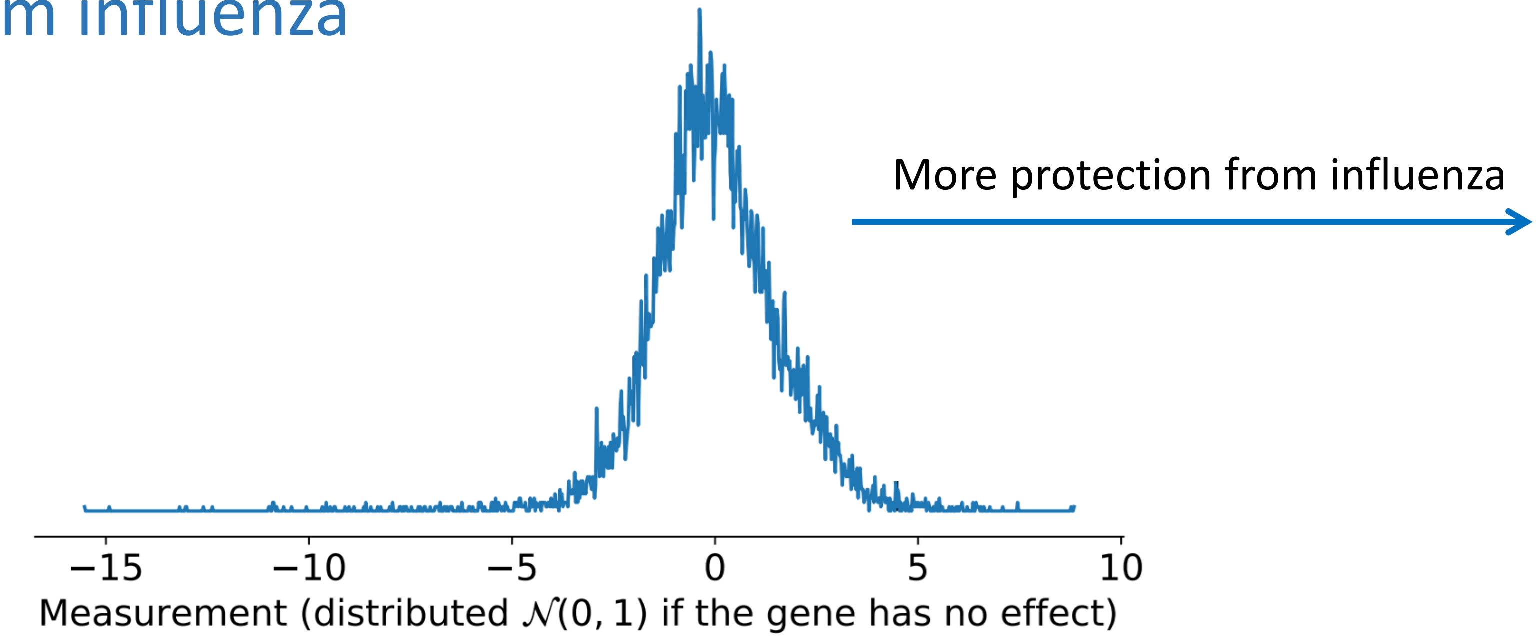
ICML 2020

UNIVERSITY *of* WASHINGTON

Example: Fruit Fly Genetics

Hao et al. (2008) measured the effect of 13,000 fruit fly genes on susceptibility to influenza

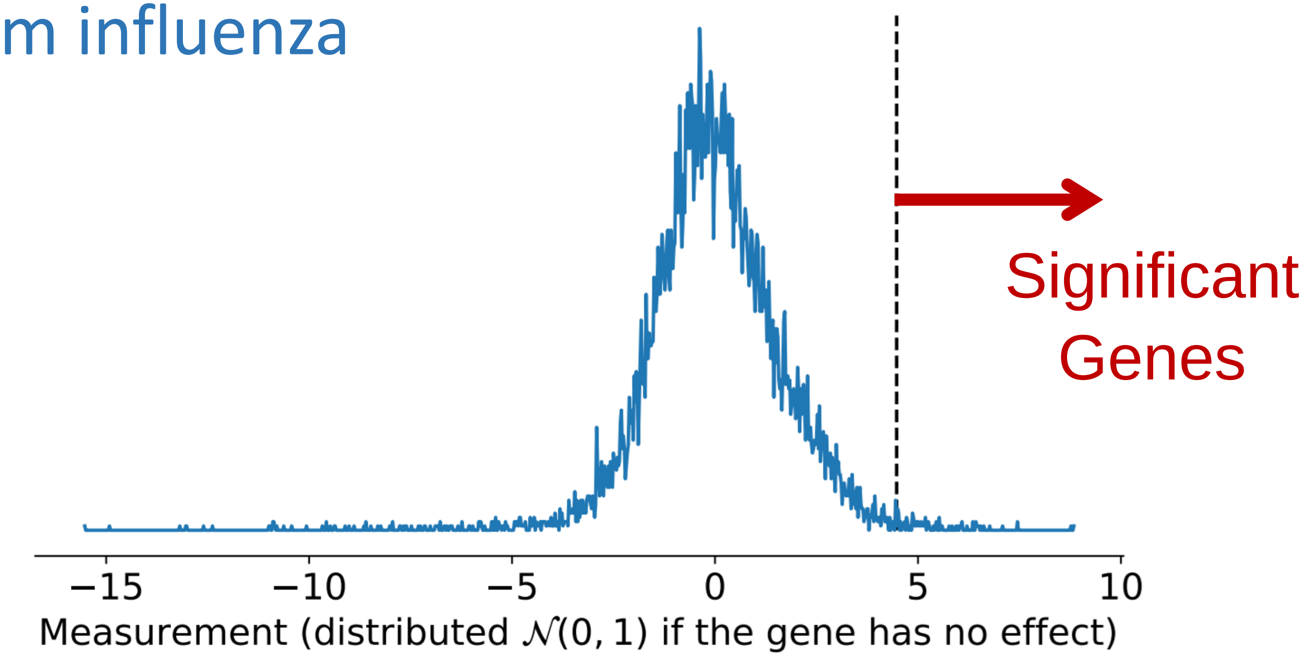
Measurements were distributed N(0,1) under the null, higher indicates protection from influenza



Example: Fruit Fly Genetics

Hao et al. (2008) measured the effect of 13,000 fruit fly genes on susceptibility to influenza

Measurements were distributed N(0,1) under the null, higher indicates protection from influenza

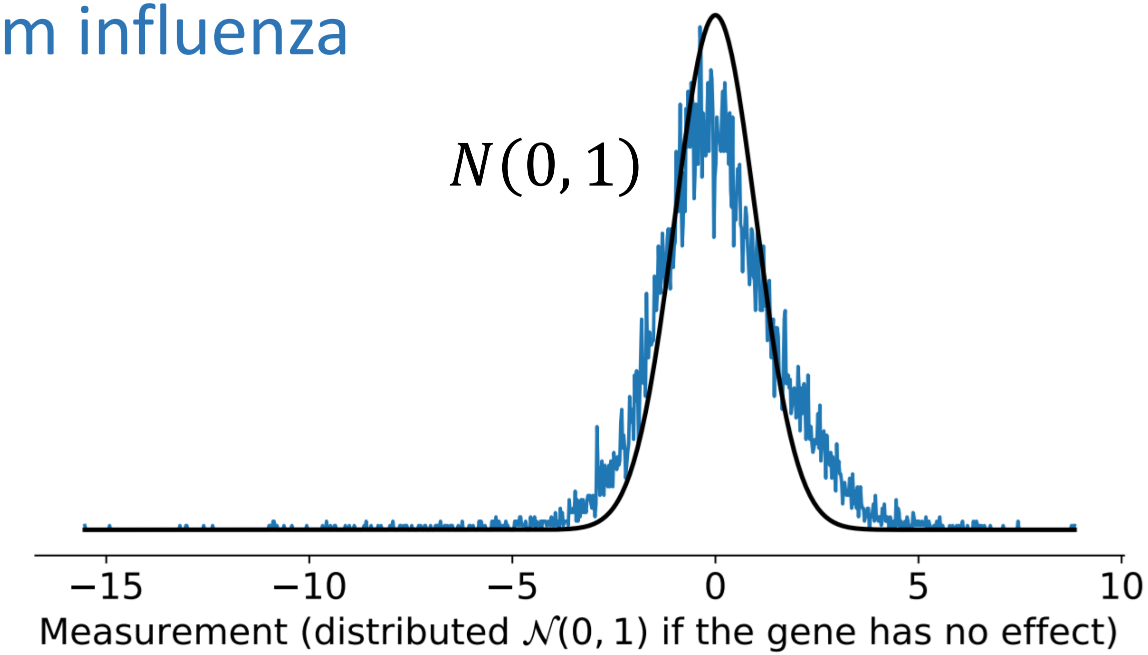


Multiple hypothesis testing identifies few discoveries

Example: Fruit Fly Genetics

Hao et al. (2008) measured the effect of 13,000 fruit fly genes on susceptibility to influenza

Measurements were distributed N(0,1) under the null, higher indicates protection from influenza

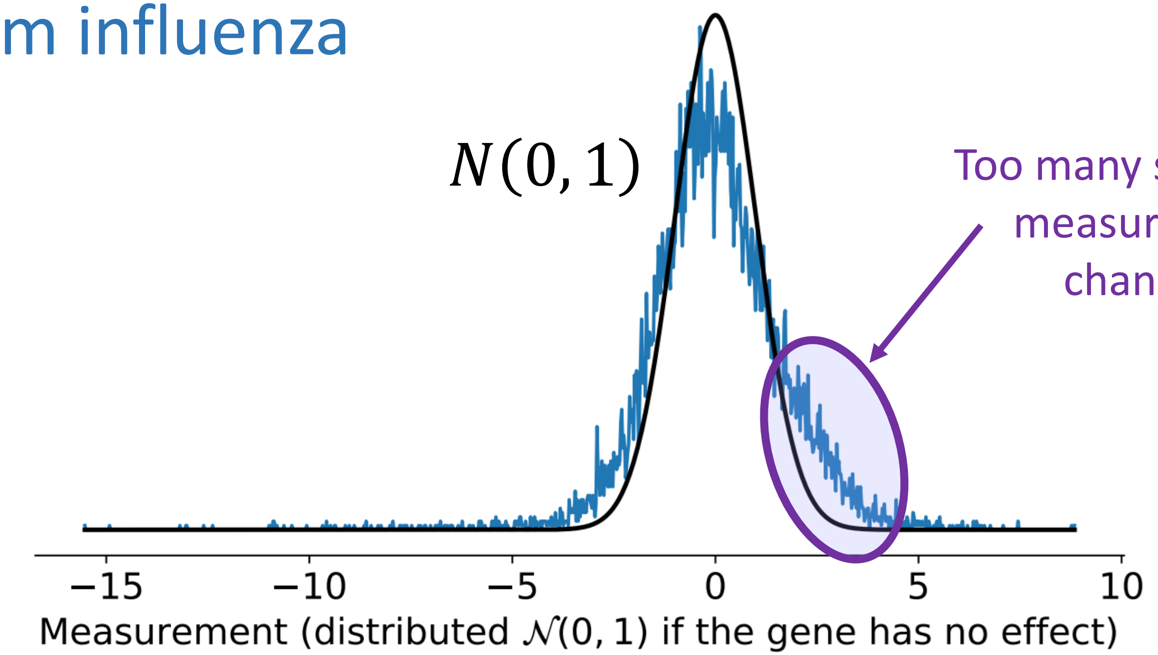


Observed distribution does not match theoretical null

Example: Fruit Fly Genetics

Hao et al. (2008) measured the effect of 13,000 fruit fly genes on susceptibility to influenza

Measurements were distributed N(0,1) under the null, higher indicates protection from influenza

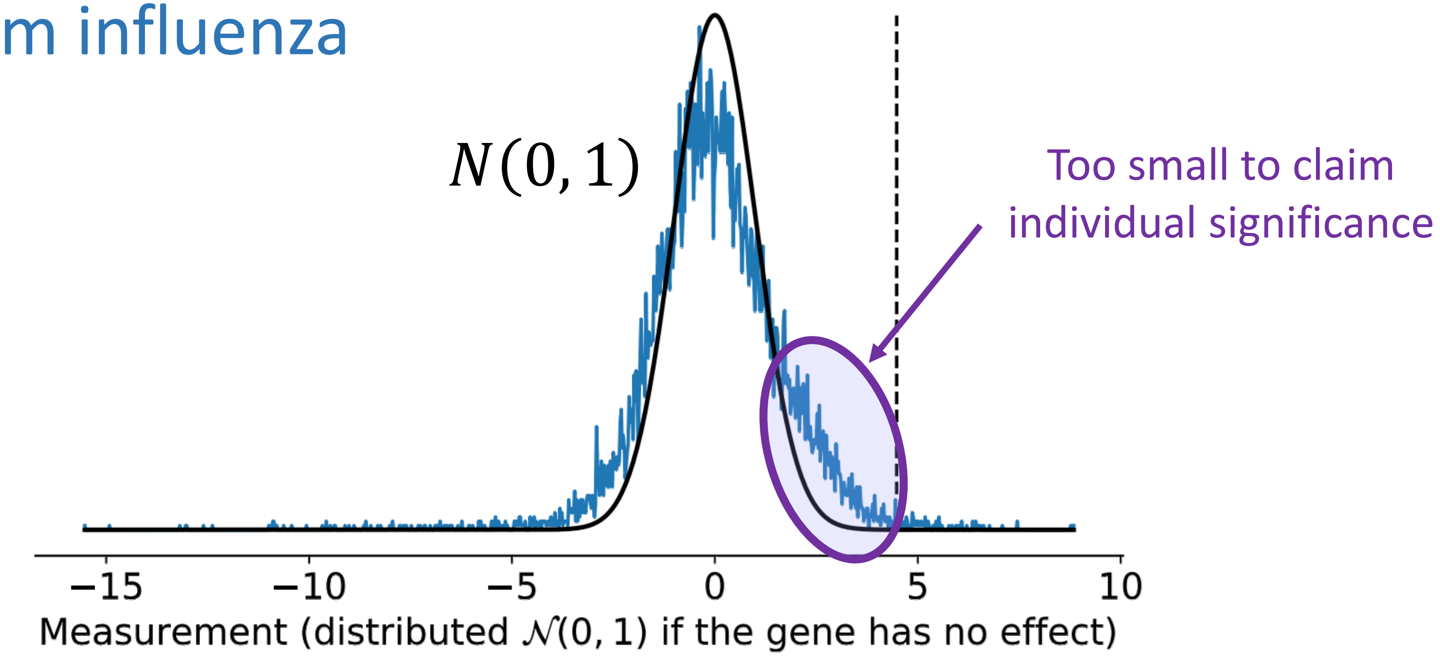


Observed distribution does not match theoretical null

Example: Fruit Fly Genetics

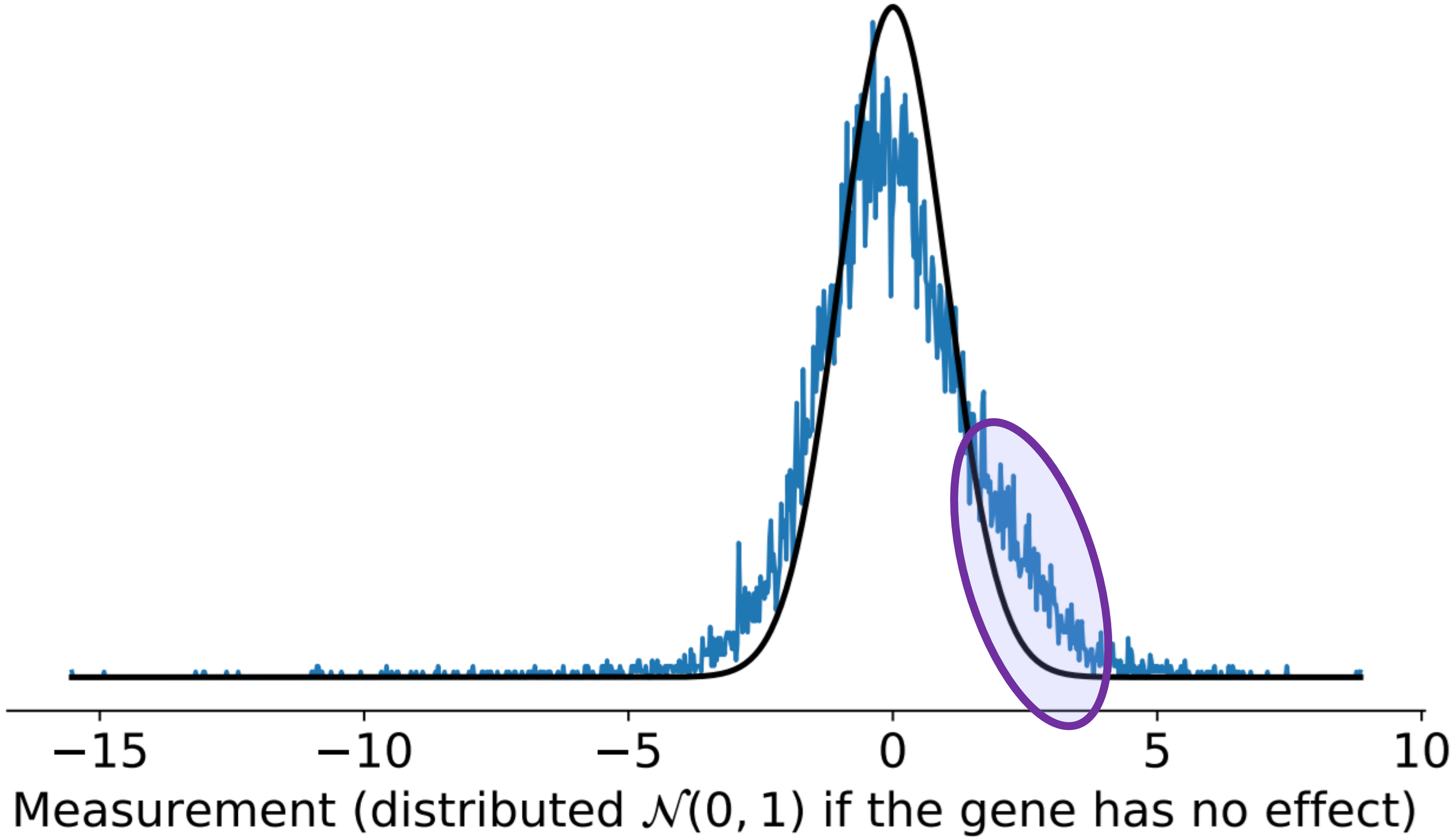
Hao et al. (2008) measured the effect of 13,000 fruit fly genes on susceptibility to influenza

Measurements were distributed N(0,1) under the null, higher indicates protection from influenza



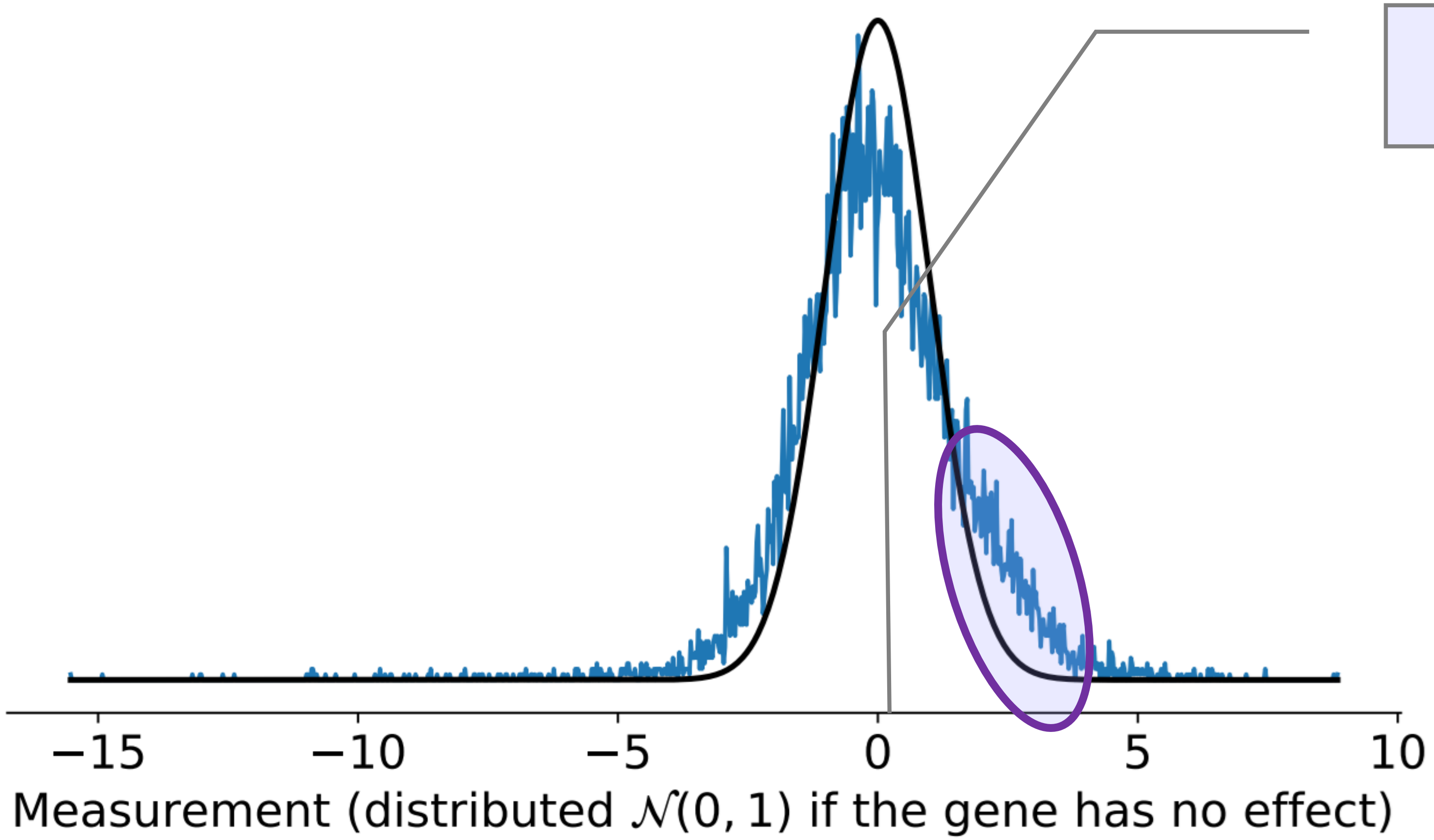
Observed distribution does not match theoretical null

Example: Fruit Fly Genetics



Idea: These genes can be **counted**,
even though they can't be **identified**

Example: Fruit Fly Genetics

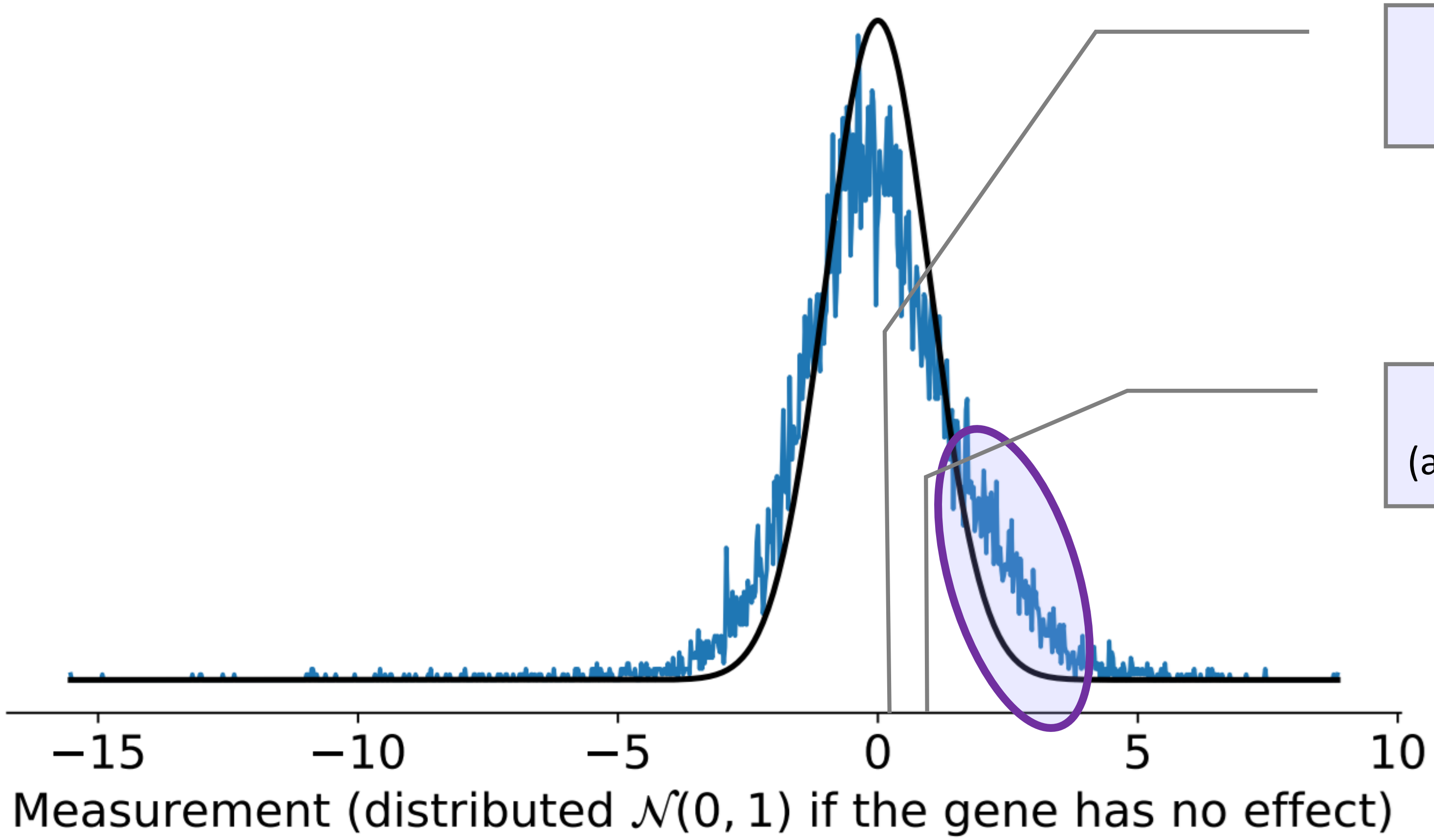


Our Estimator

>7% of genes have effect size >1/4
(at least **8% increase in influenza resistance**)

Idea: These genes can be **counted**,
even though they can't be **identified**

Example: Fruit Fly Genetics



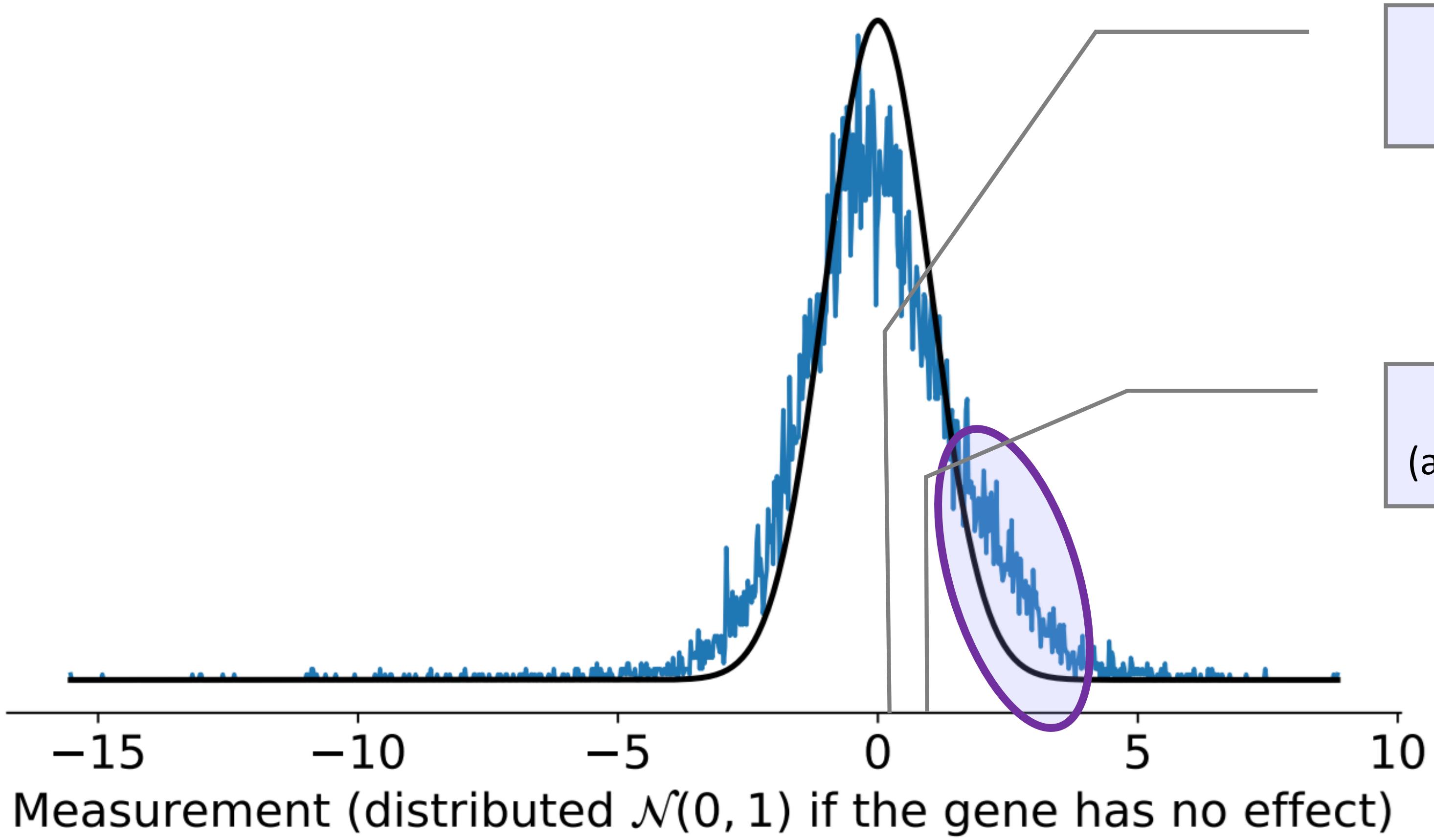
Our Estimator

>7% of genes have effect size >1/4
(at least **8% increase in influenza resistance**)

>2% of genes have effect size >1
(at least **28% increase in influenza resistance**)

Idea: These genes can be **counted**,
even though they can't be **identified**

Example: Fruit Fly Genetics



Our Estimator

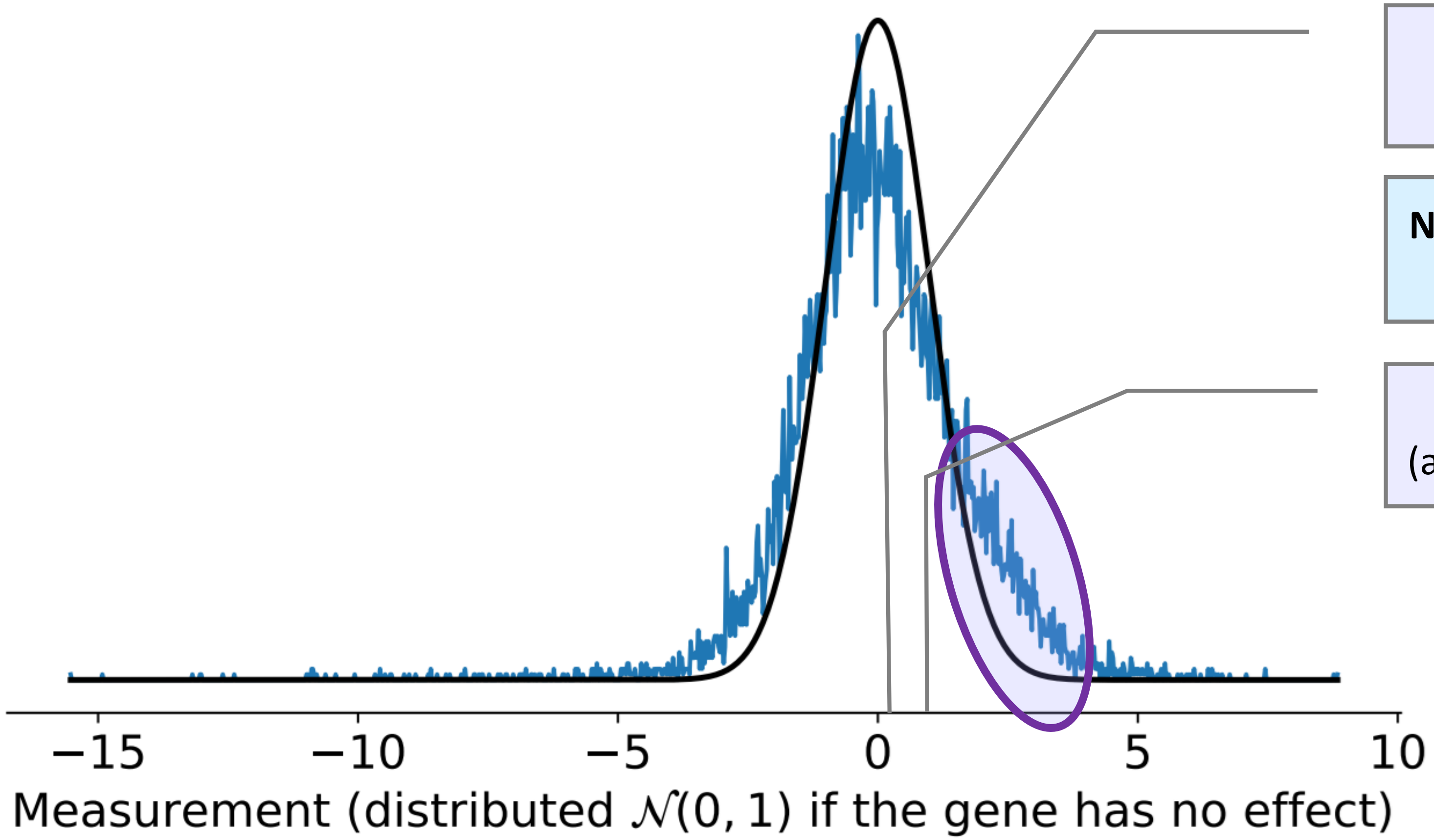
>7% of genes have effect size >1/4
(at least **8% increase in influenza resistance**)

>2% of genes have effect size >1
(at least **28% increase in influenza resistance**)

Idea: These genes can be **counted**,
even though they can't be **identified**

Enables **power analysis** for
future **experimental designs**

Example: Fruit Fly Genetics



Our Estimator

>7% of genes have effect size >1/4
(at least **8% increase in influenza resistance**)

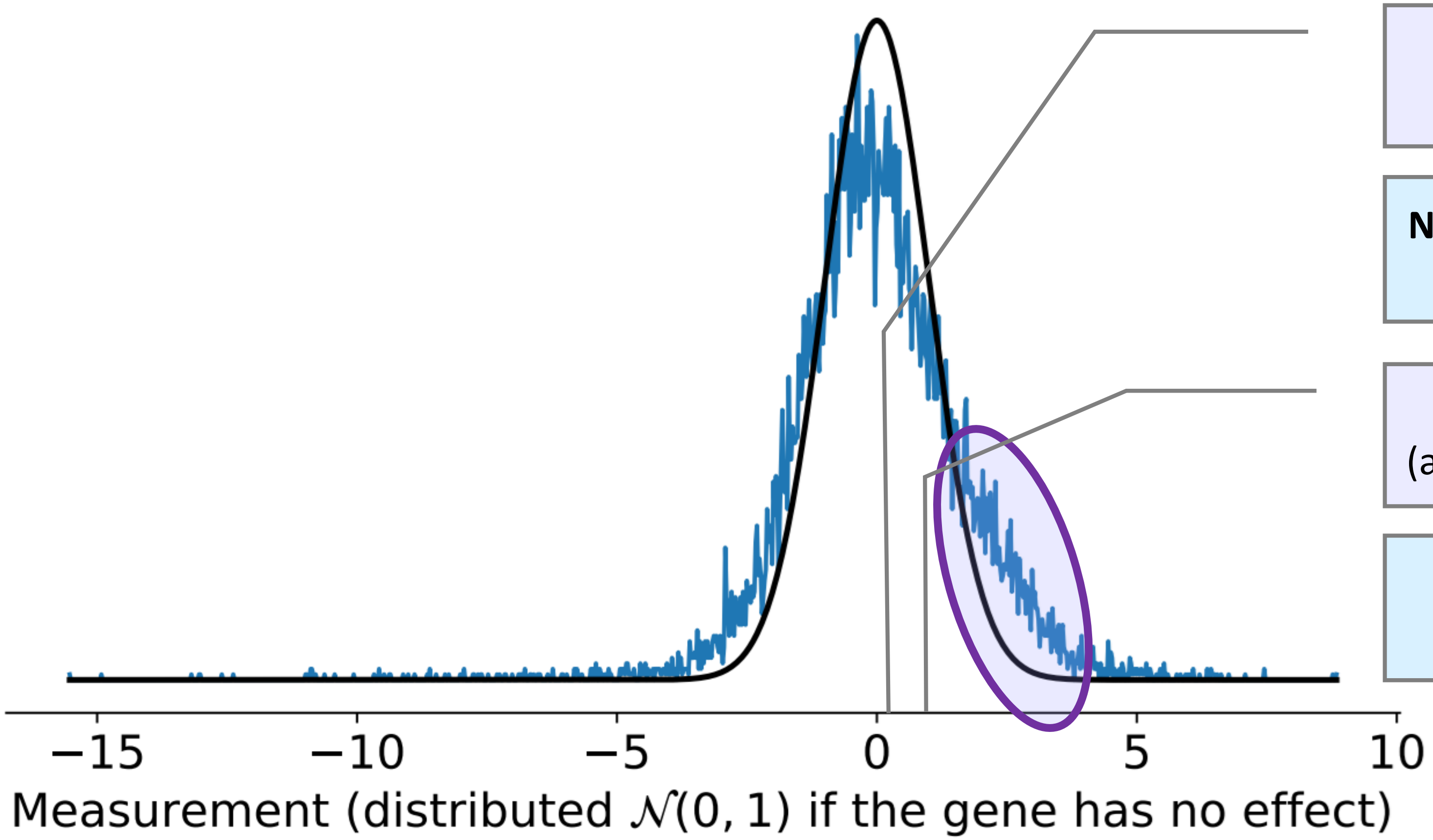
Next Experiment: Take precise measurements (e.g., use many replications) to **identify** these genes

>2% of genes have effect size >1
(at least **28% increase in influenza resistance**)

Idea: These genes can be **counted**, even though they can't be **identified**

Enables **power analysis** for future **experimental designs**

Example: Fruit Fly Genetics



Our Estimator

>7% of genes have effect size >1/4
(at least **8% increase in influenza resistance**)

Next Experiment: Take precise measurements (e.g., use many replications) to **identify** these genes

>2% of genes have effect size >1
(at least **28% increase in influenza resistance**)

Next Experiment: Take less precise measurements, **identify** fewer genes

Idea: These genes can be **counted**,
even though they can't be **identified**

Enables **power analysis** for
future **experimental designs**

Formal problem statement

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$

μ_i is the (unknown) effect size

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \dots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

μ_i is the (unknown) effect size

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \dots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

μ_i is the (unknown) effect size

E.g. $f(\mu_i) = N(\mu_i, 1)$

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$ — μ_i is the (unknown) effect size

Observe $X_i \sim f(\mu_i)$ — E.g. $f(\mu_i) = N(\mu_i, 1)$

Identification: Which $\mu_i > 0$?

Counting: What is the probability $P_{\mu \sim \nu_*}(\mu > 0)$?

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

μ_i is the (unknown) effect size

E.g. $f(\mu_i) = N(\mu_i, 1)$

Identification: Which $\mu_i > 0$?

Counting: What is the probability $P_{\mu \sim \nu_*}(\mu > \gamma)$, for all γ?

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \dots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

μ_i is the (unknown) effect size

E.g. $f(\mu_i) = N(\mu_i, 1)$

Identification: Which $\mu_i > 0$? (Returns a set in [n])

Counting: What is the probability $P_{\mu\sim\nu_*}(\mu > \gamma)$, for all γ?

(Returns a fraction)

Formal problem statement

We view multiple hypothesis testing from the perspective of **learning mixture distributions**

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$ — μ_i is the (unknown) effect size

Observe $X_i \sim f(\mu_i)$ — E.g. $f(\mu_i) = N(\mu_i, 1)$

Goal

Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu \sim \nu_*}(\mu > \gamma)$, for all γ

Constraint Never overestimate the true fraction

Related work

Estimating the number of non-nulls ($\mu \neq 0$)

- Early techniques [Schweder and Spjøtvoll, 1982; Genovese et al., 2004; Meinshausen et al., 2006] relied on **uniformity of p-values** under the null
- Techniques do not extend to **arbitrary thresholds** ("How many genes improved influenza resistance by at least 20%?")

Related work

Estimating the number of non-nulls ($\mu \neq 0$)

- Early techniques [Schweder and Spjøtvoll, 1982; Genovese et al., 2004; Meinshausen et al., 2006] relied on **uniformity of p-values** under the null
- Techniques do not extend to **arbitrary thresholds** ("How many genes improved influenza resistance by at least 20%?")

Plug-in estimators

- **Estimate the entire density** ν, then compute $P_\nu(\mu > \gamma)$
- Does not respect our **constraint**, that we cannot overestimate

Related work

Estimating the number of non-nulls ($\mu \neq 0$)

- Early techniques [Schweder and Spjøtvoll, 1982; Genovese et al., 2004; Meinshausen et al., 2006] relied on **uniformity of p-values** under the null
- Techniques do not extend to **arbitrary thresholds** (“How many genes improved influenza resistance by at least 20%?”)

Plug-in estimators

- **Estimate the entire density** ν, then compute $P_\nu(\mu > \gamma)$
- Does not respect our **constraint**, that we cannot overestimate

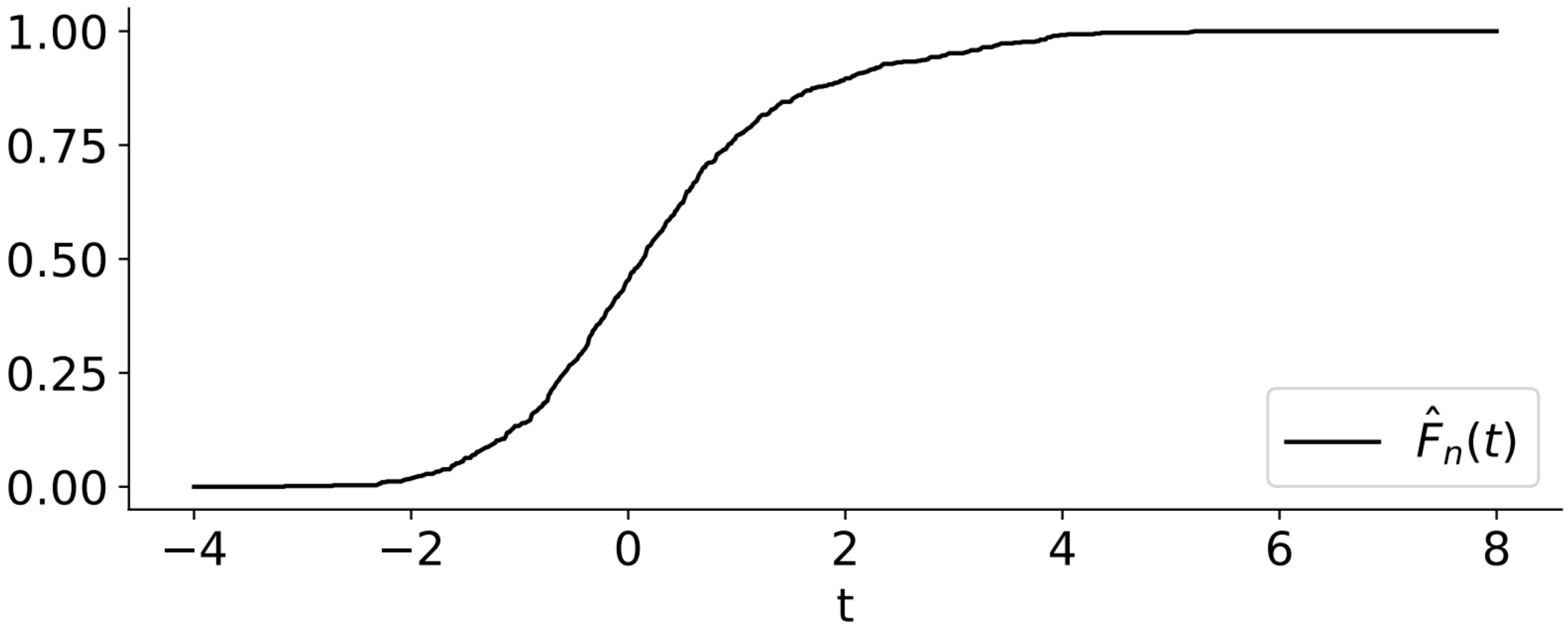
Connections to False Discovery Rate (FDR) control

- **Tighter FDR control** can be obtained by knowing number of non-nulls
- Previous methods either do not satisfy our **constraint** [Storey, 2002; Li and Barber, 2019], or perform poorly in our **regime of interest** (many hypotheses, small effect sizes) [Stephens, 2016; Katsevich and Ramdas, 2018]

Our Estimator

Our Estimator

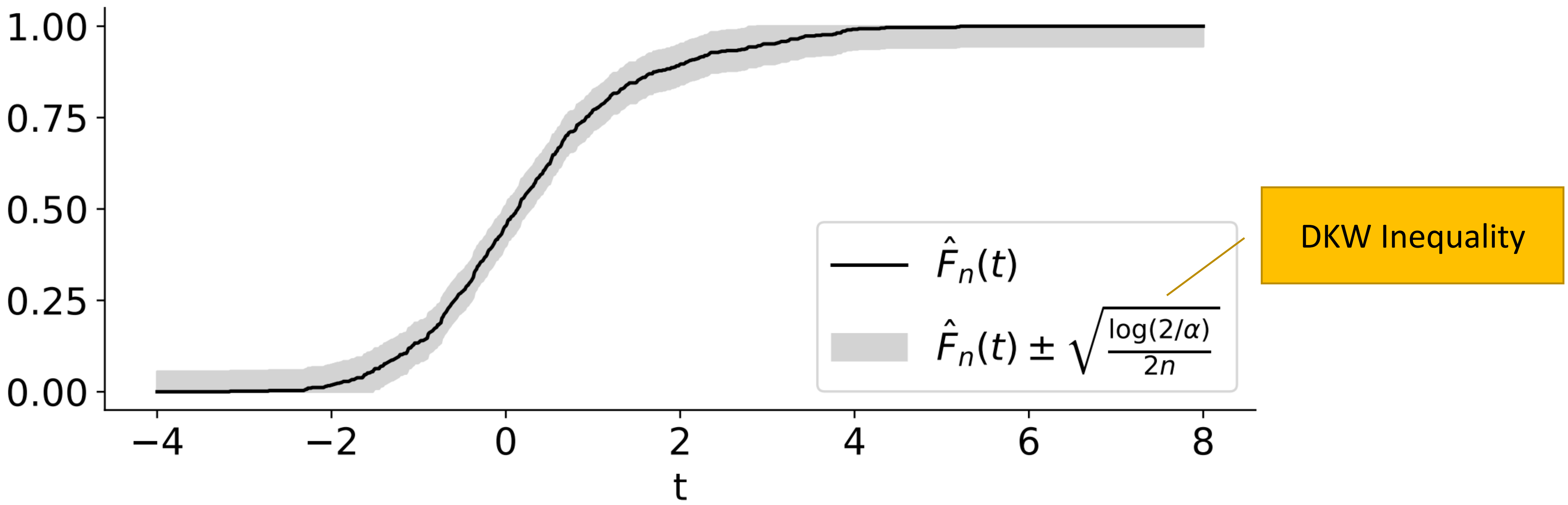
Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



Step 1 Consider the **empirical CDF** (Cumulative Distribution Function)

Our Estimator

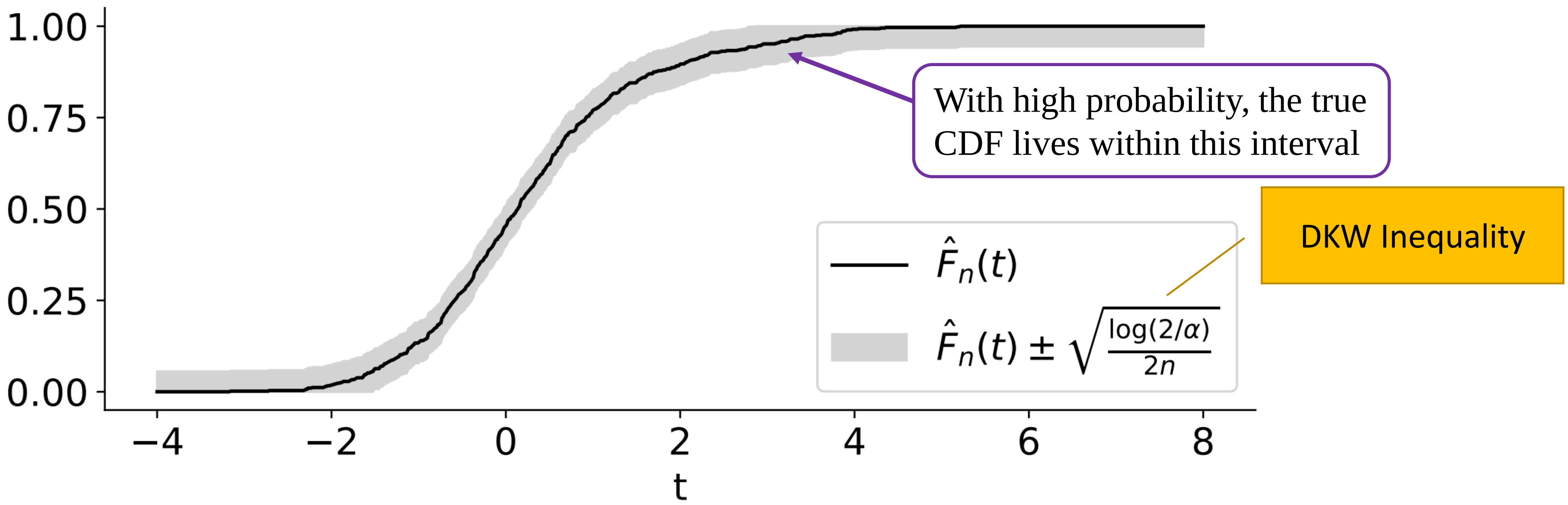
Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate **confidence intervals** on the true CDF

Our Estimator

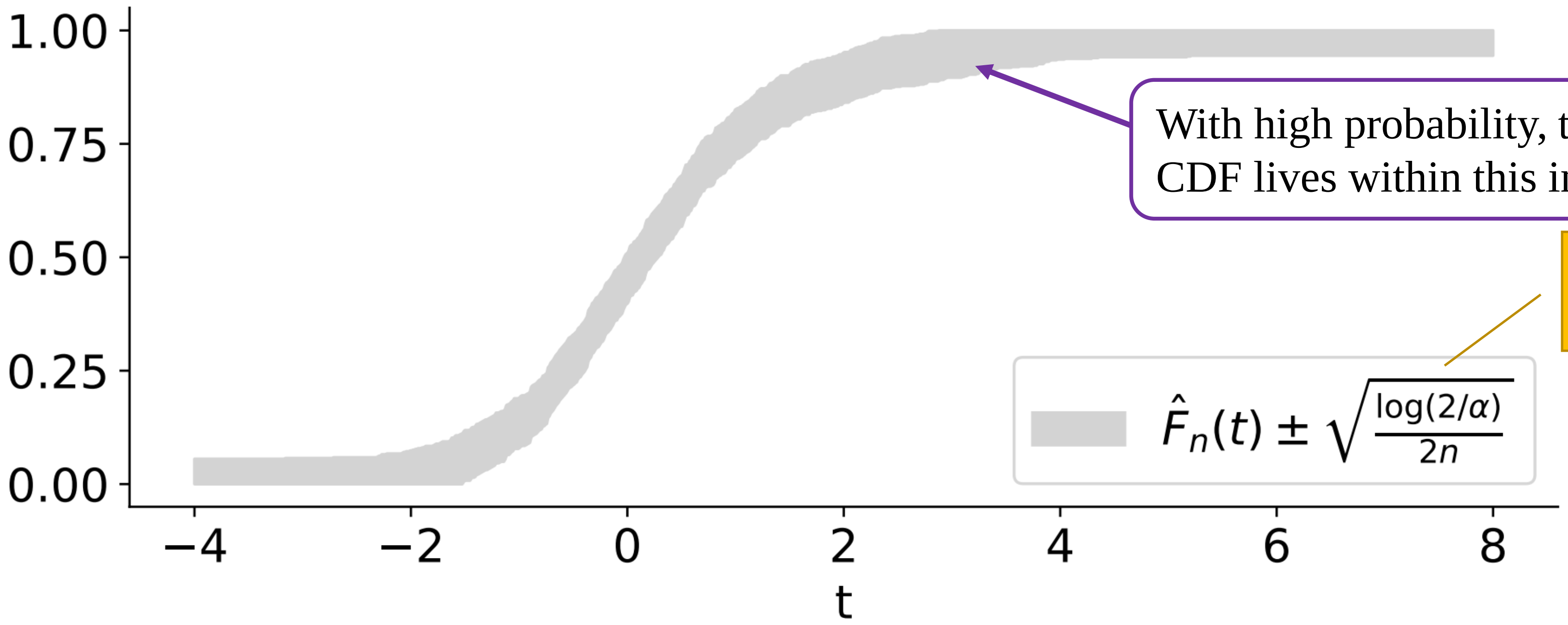
Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate **confidence intervals** on the true CDF

Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



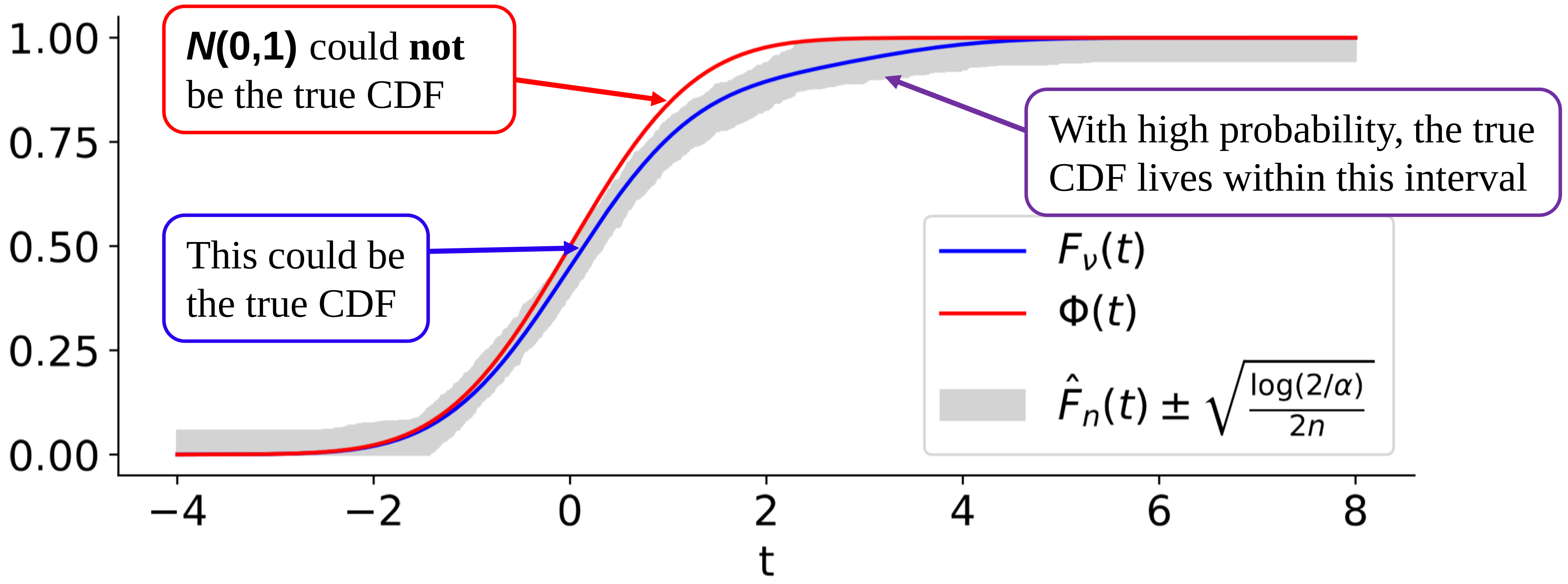
With high probability, the true CDF lives within this interval

DKW Inequality

Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate **confidence intervals** on the true CDF

Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate

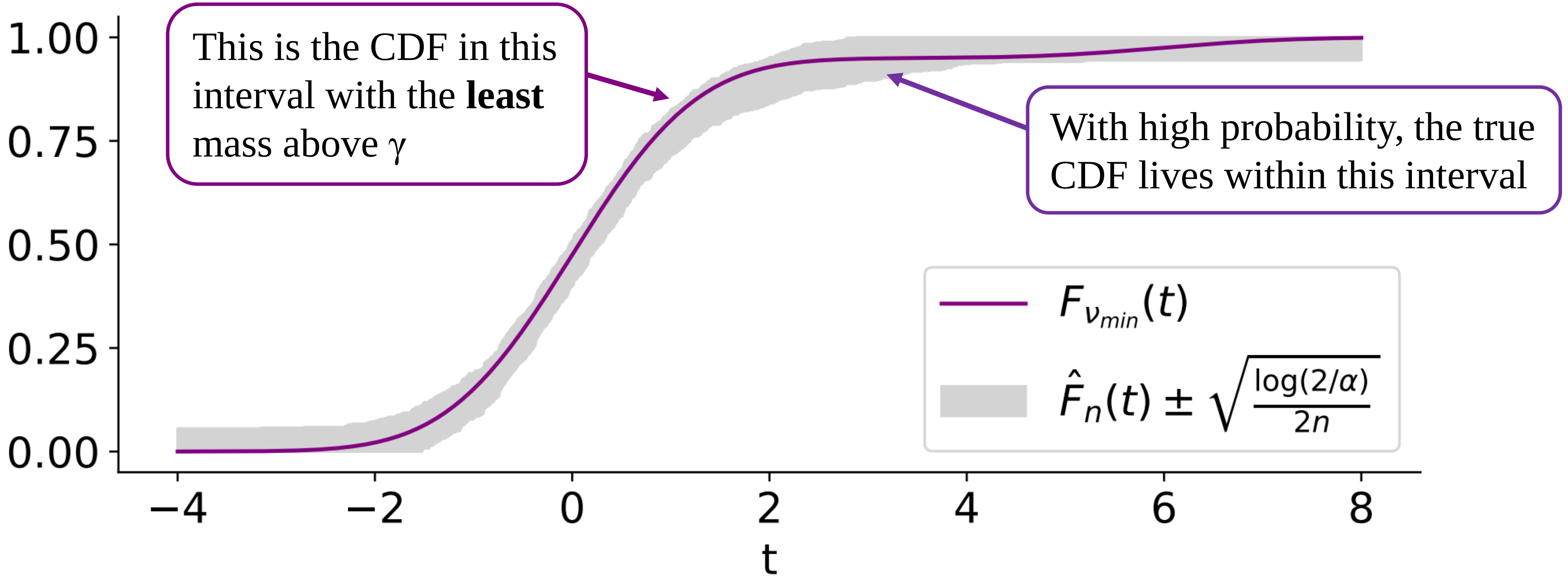


Step 1 Consider the empirical CDF (Cumulative Distribution Function)

Step 2 Generate **confidence intervals** on the true CDF

Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate confidence intervals on the true CDF
Step 3 Return the **smallest amount of mass** that could **feasibly have generated** the empirical CDF

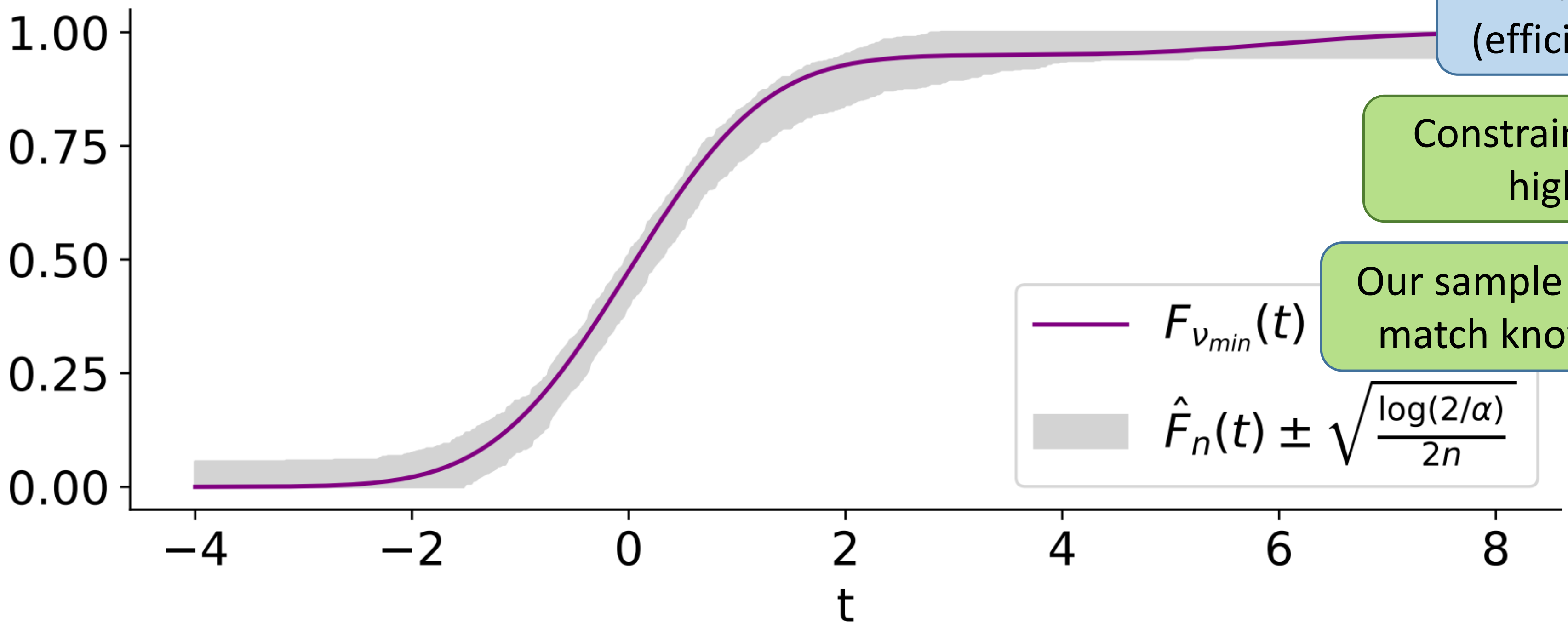
Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate

A convex program! (efficiently computable)

Constraint is satisfied with high probability

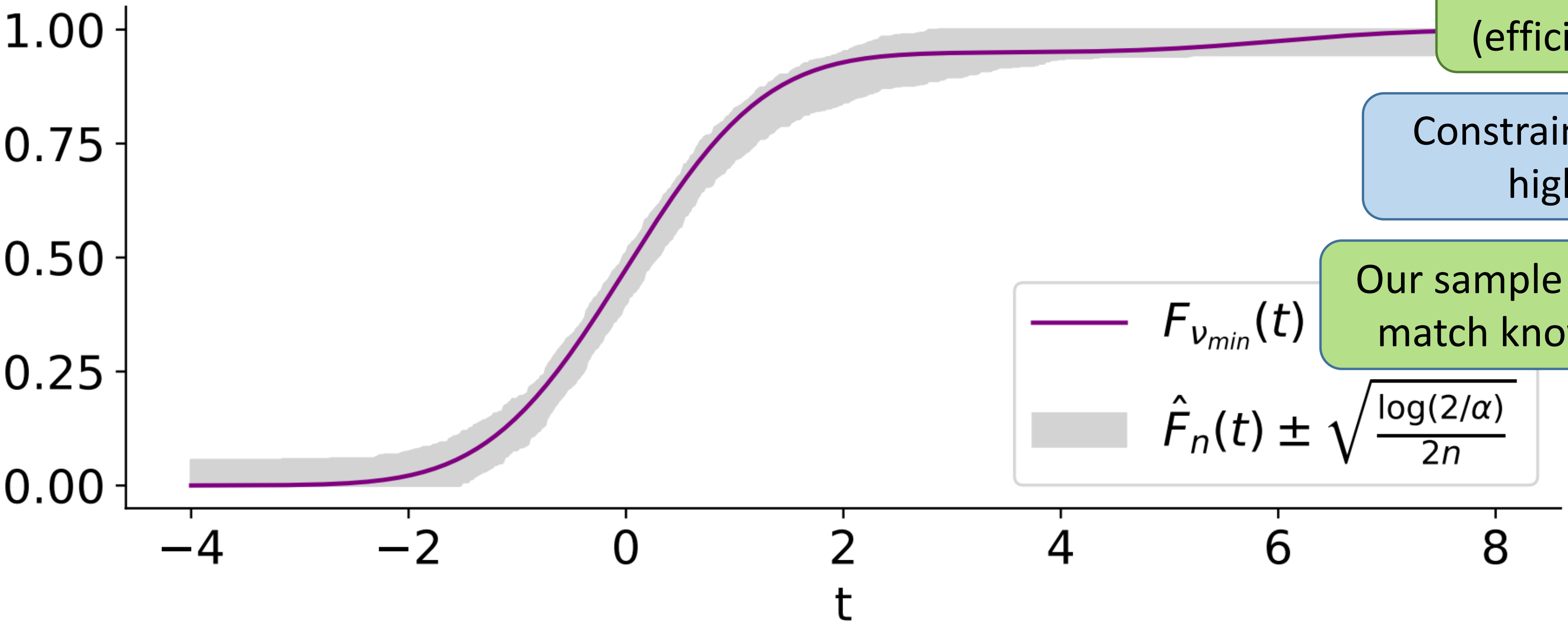
Our sample complexity results match known lower bounds



Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate confidence intervals on the true CDF
Step 3 Return the **smallest amount of mass** that could **feasibly have generated** the empirical CDF

Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



A convex program! (efficiently computable)

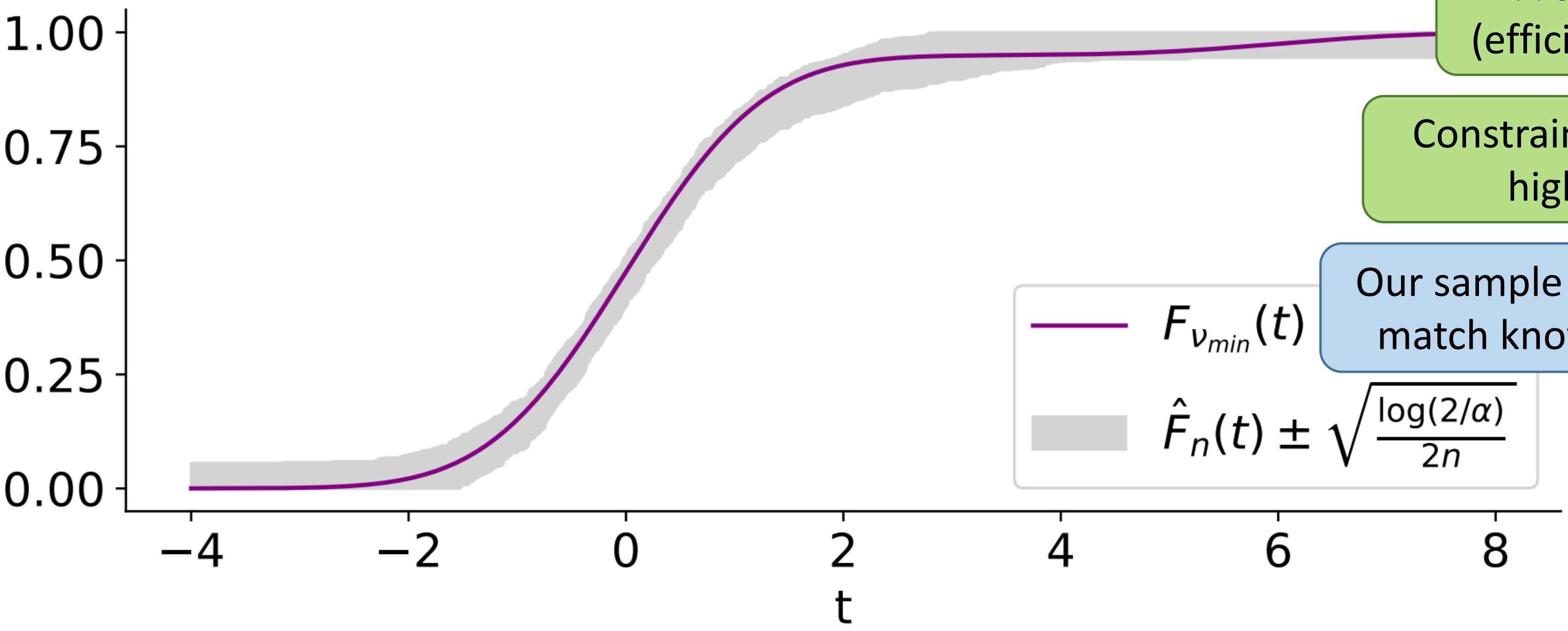
Constraint is satisfied with high probability

Our sample complexity results match known lower bounds

Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate confidence intervals on the true CDF
Step 3 Return the **smallest amount of mass** that could **feasibly have generated** the empirical CDF

Our Estimator

Goal Estimate $\zeta_{\nu_*}(\gamma) := \mathbb{P}_{\nu_*}(\mu_i > \gamma)$
Constraint Never overestimate



A convex program! (efficiently computable)

Constraint is satisfied with high probability

Our sample complexity results match known lower bounds

Step 1 Consider the empirical CDF (Cumulative Distribution Function)
Step 2 Generate confidence intervals on the true CDF
Step 3 Return the **smallest amount of mass** that could **feasibly have generated** the empirical CDF

Theorem

Our estimator provides the following guarantees:

With probability $1 - \alpha$, **does not overestimate** $\zeta_{\nu_*}(\gamma)$ for any γ

For $i = 1, 2, \dots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu \sim \nu_*}(\mu > \gamma)$

Theorem

Our estimator provides the following guarantees:

With probability $1 - \alpha$, **does not overestimate** $\zeta_{\nu_*}(\gamma)$ for any γ

With probability $1 - \delta$, estimate is **at most ε from the truth** whenever

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

For $i = 1, 2, \ldots, n$
    Draw $\mu_i \sim \nu_*$
    Observe $X_i \sim f(\mu_i)$
Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu\sim\nu_*}(\mu > \gamma)$

Theorem

Our estimator provides the following guarantees:

With probability $1-\alpha$, **does not overestimate** $\zeta_{\nu_*}(\gamma)$ for any γ

With probability $1-\delta$, estimate is **at most ε from the truth** whenever

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

F_{ν_*}

CDFs F_ν corresponding to all mixing distributions ν with less than $\zeta_{\nu_*}(\gamma)-\varepsilon$ probability mass above γ

For $i = 1, 2, \dots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu\sim\nu_*}(\mu > \gamma)$

Theorem

Our estimator provides the following guarantees:

With probability $1-\alpha$, **does not overestimate** $\zeta_{\nu_*}(\gamma)$ for any γ

With probability $1-\delta$, estimate is **at most ε from the truth** whenever

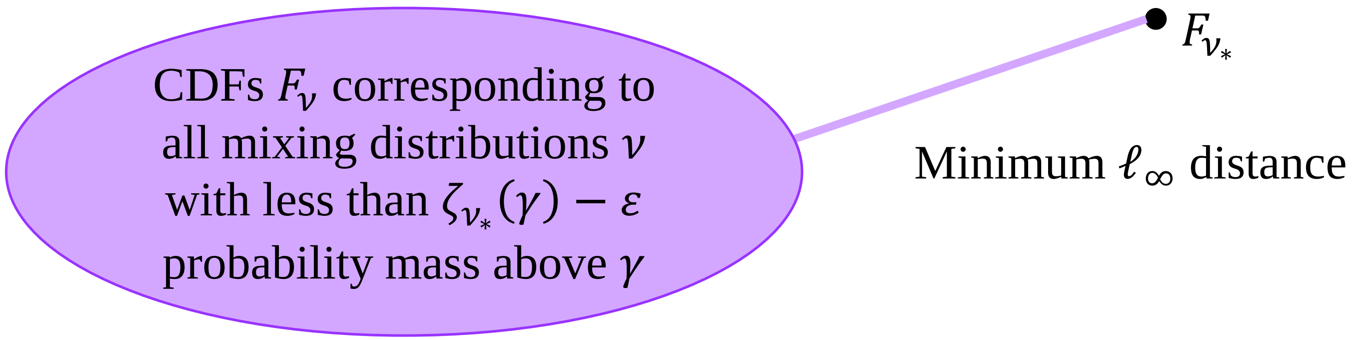
$$n \geq \frac{\log\left(\frac{4}{\alpha\delta}\right)}{\left(\min_{\nu:\ \mathbb{P}_{\mu\sim\nu}(\mu>\gamma)\leq\zeta_{\nu_*}(\gamma)-\varepsilon} ||F_\nu - F_{\nu_*}||_\infty\right)^2}.$$

For $i = 1, 2, \ldots, n$

Draw $\mu_i \sim \nu_*$

Observe $X_i \sim f(\mu_i)$

Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu\sim\nu_*}(\mu > \gamma)$



Theorem

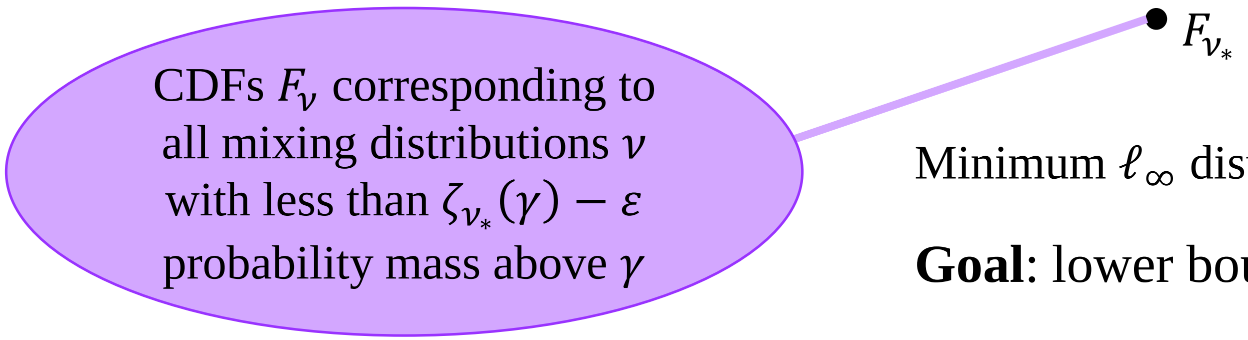
Our estimator provides the following guarantees:

With probability $1-\alpha$, **does not overestimate** $\zeta_{\nu_*}(\gamma)$ for any γ

With probability $1-\delta$, estimate is **at most ε from the truth** whenever

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

For $i = 1, 2, \ldots, n$
  Draw $\mu_i \sim \nu_*$
  Observe $X_i \sim f(\mu_i)$
Estimate $\zeta_{\nu_*}(\gamma) = P_{\mu\sim\nu_*}(\mu > \gamma)$



Minimum ℓ_∞ distance

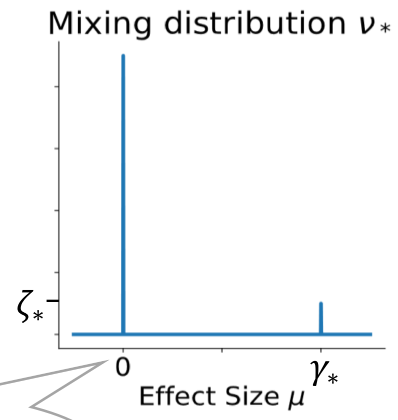
Goal: lower bound this distance

Counting at least half of the discoveries

Let $X_i \sim N(\mu_i, 1)$ be drawn from a **mixture of Gaussians**, with ζ_* alternate hypotheses of effect size $\gamma_* < 1$

With probability at least $1 - \delta$, our estimator **detects over half** of the alternate hypotheses (i.e., $\hat{\zeta}_n(0) > \frac{1}{2}\zeta_*$), whenever

$$n \gtrsim \frac{\log\left(\frac{2}{\delta}\right)}{\zeta_*^2 \gamma_*^4}.$$

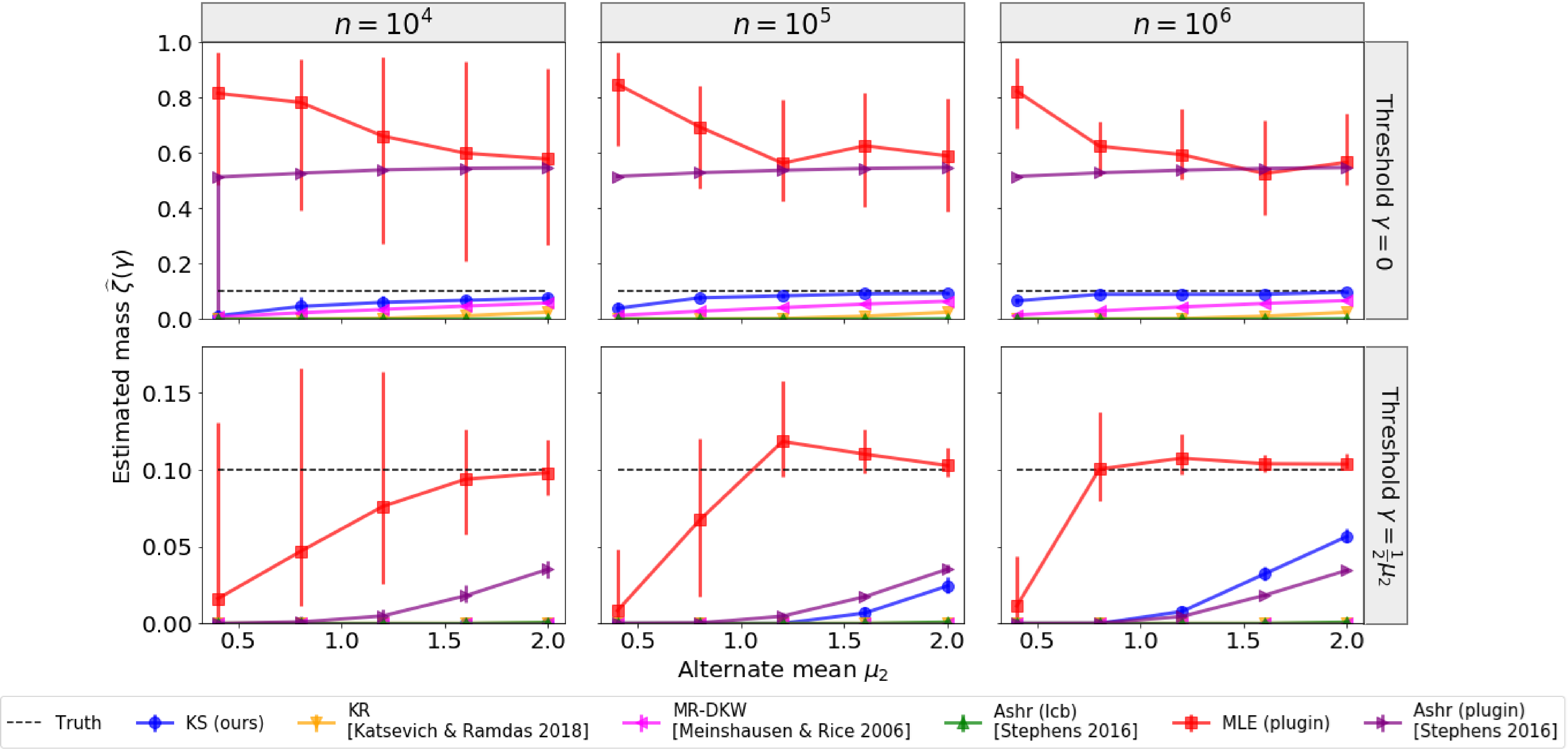


How much mass is (strictly) above 0?

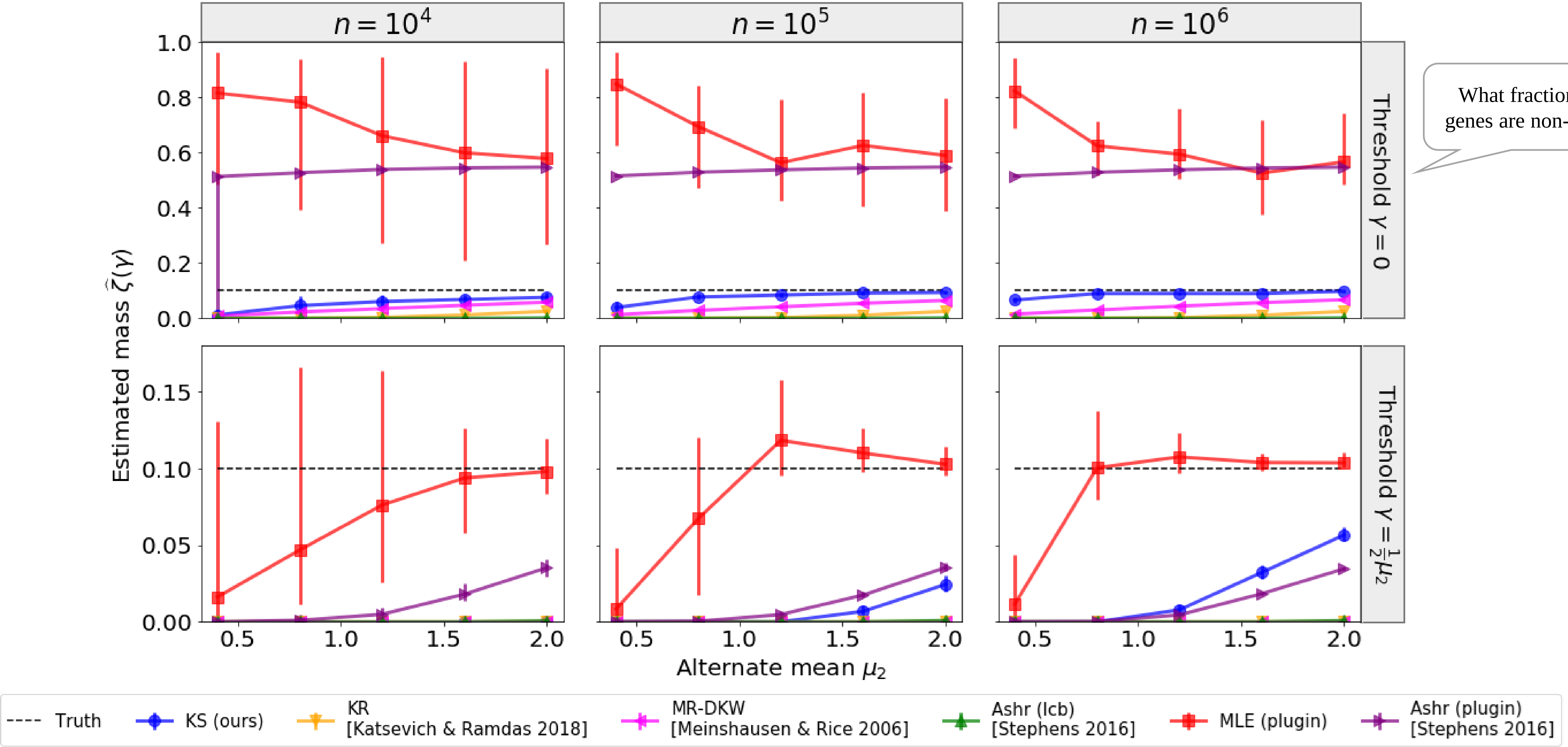
Matches a novel lower bound

Experiments

Comparisons to baselines

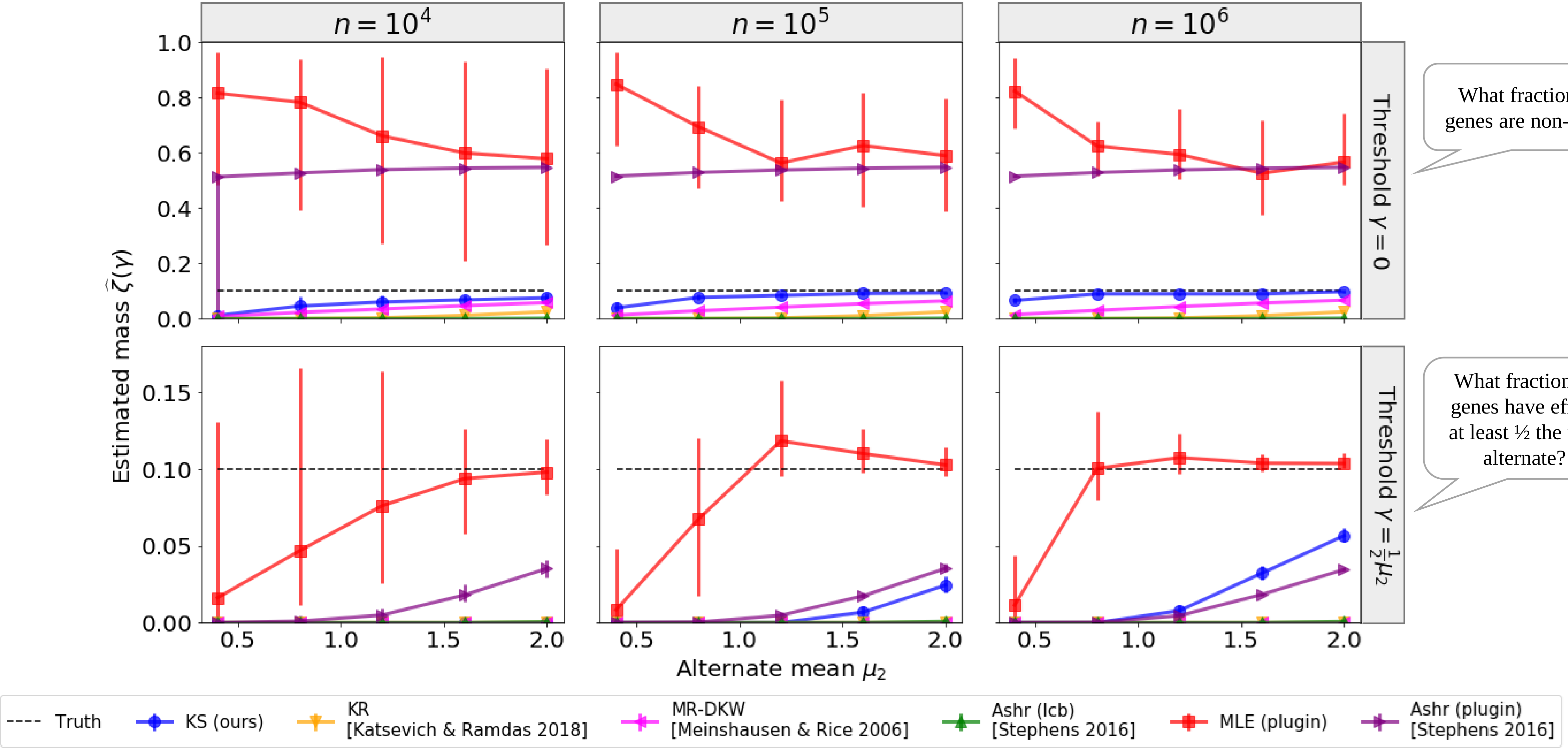


Comparisons to baselines



What fraction of genes are non-null?

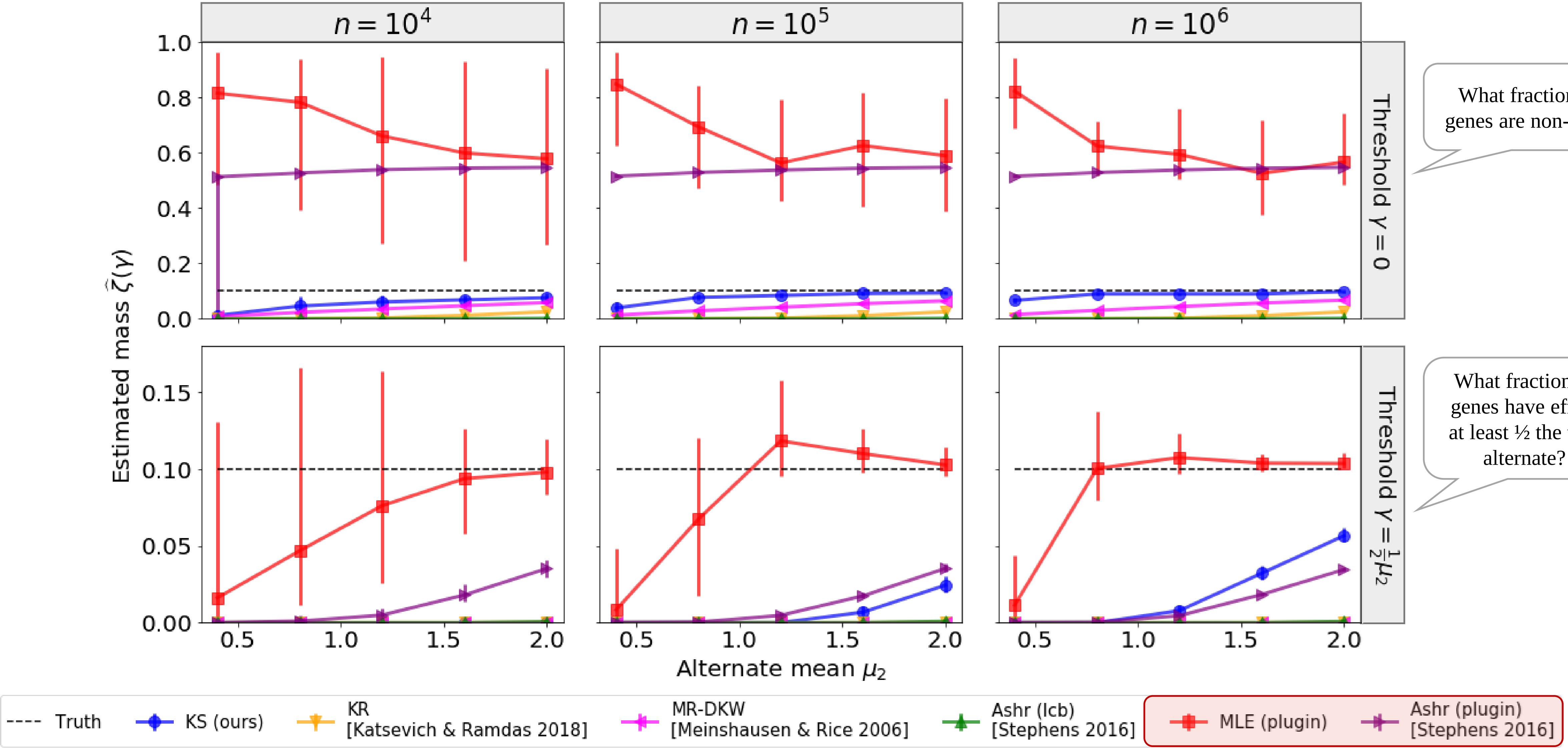
Comparisons to baselines



What fraction of genes are non-null?

What fraction of genes have effect at least ½ the true alternate?

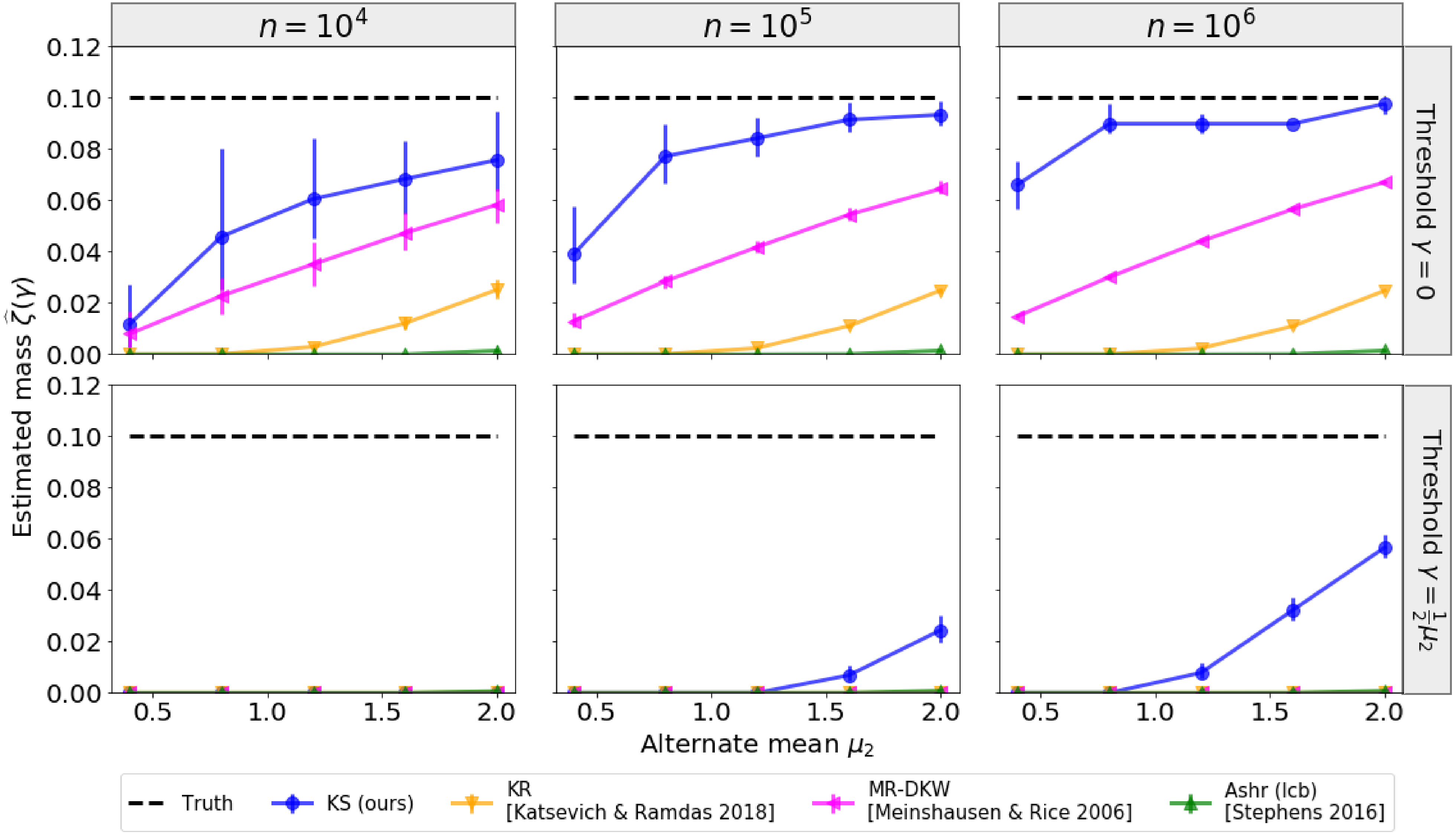
Comparisons to baselines



What fraction of genes are non-null?

What fraction of genes have effect at least ½ the true alternate?

Comparisons to baselines



What fraction of genes are non-null?

What fraction of genes have effect at least ½ the true alternate?

Other applications of this estimator

Standardized Testing

Each school has some μ_i indicating its students' true performance

We observe X_i, a noisy measurement of μ_i (e.g., students' average exam score)

Our estimator: "at least Y% of schools are below proficient in math"

Interesting on its own, or to suggest further testing to identify these schools

Other applications of this estimator

Standardized Testing

Each school has some μ_i indicating its students' true performance

We observe X_i, a noisy measurement of μ_i (e.g., students' average exam score)

Our estimator: "at least Y% of schools are below proficient in math"

Interesting on its own, or to suggest further testing to identify these schools

Public Health*

Each person has some μ_i indicating their susceptibility to the flu (variable due to age, health, etc.)

We observe X_i, the number of flu seasons they were sick, in the past five years

Our estimator: "at most Y% of people have a 25% chance or greater of getting sick in a given year" (Impossible to identify these people with confidence)

*Example due to Tian, Kong and Valiant (2017)