# A Novel Cost-Effective Surveillance Strategy that Guarantees Differential Privacy

Ahmer Raza  Christopher McMahan  Rafael D'Oliveira

School of Mathematical and Statistical Sciences
Clemson University

April 18, 2025

# Overview

**Goal**: Estimate prevalence of a binary characteristic (e.g., disease status).

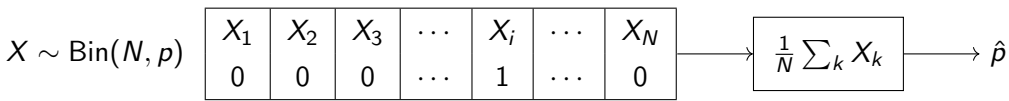
- Characteristic is low-prevalence
- Want approach to be cost-effective and inherently private

# Overview

**Goal**: Estimate prevalence of a binary characteristic (e.g., disease status).

- Characteristic is low-prevalence
- Want approach to be cost-effective and inherently private

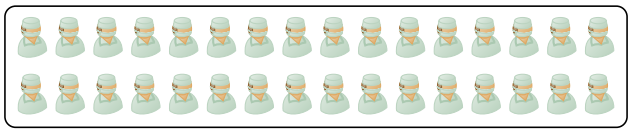
**Usual Method**:




- Expensive for large $N$
- No inherent privacy
- Test error results in a biased estimator

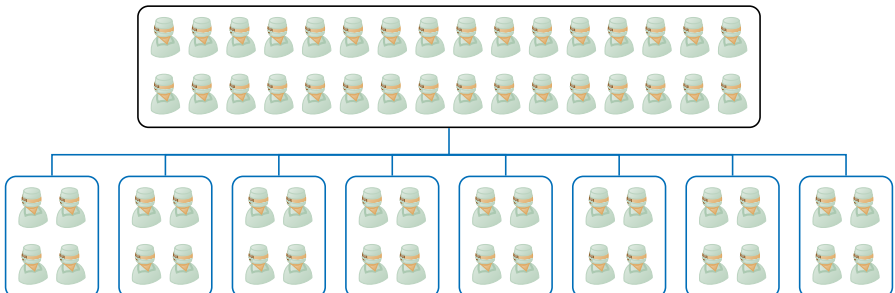
# Overview

**Solution**: Divide $N$ individuals into $J$ groups of size $c$, and test pooled samples $Z_j$.



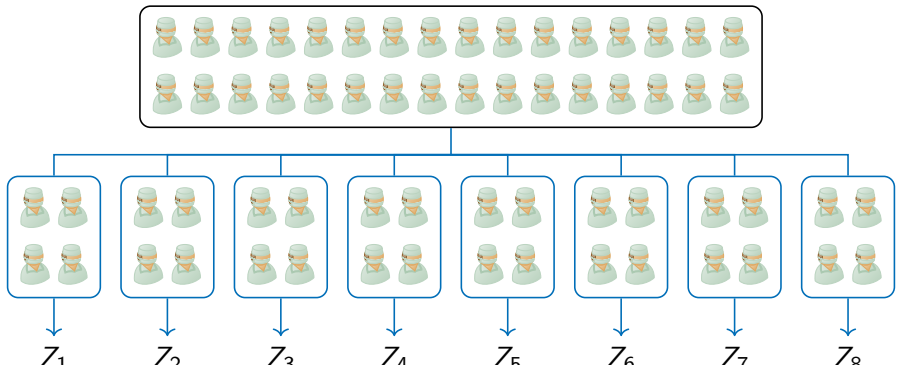
# Overview

**Solution**: Divide $N$ individuals into $J$ groups of size $c$, and test pooled samples $Z_j$.



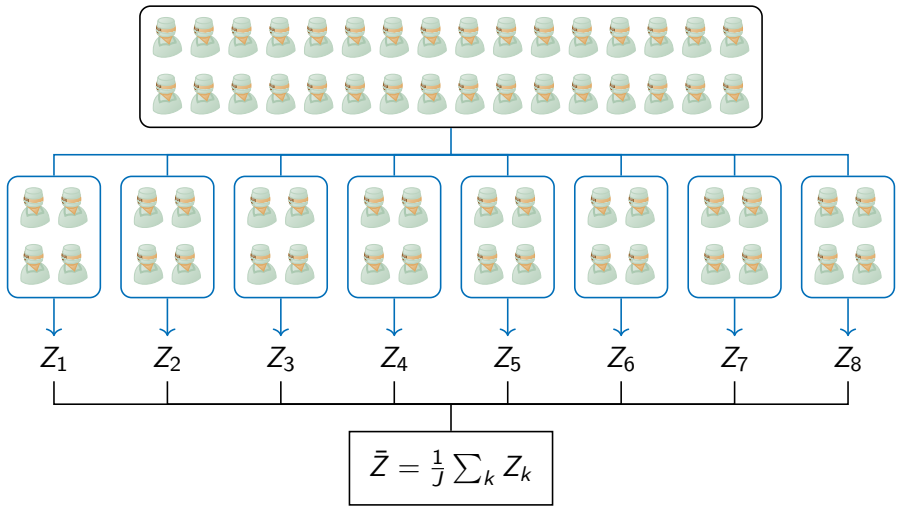
# Overview

**Solution**: Divide $N$ individuals into $J$ groups of size $c$, and test pooled samples $Z_j$.



# Overview

**Solution**: Divide $N$ individuals into $J$ groups of size $c$, and test pooled samples $Z_j$.



# Estimation and Optimality

## Theorem (Maximum Likelihood Estimator)

Let $r = S_p + S_e - 1$ and $\pi = \Pr[Z_j = 1] = S_e + r(1-p)^c$. Then the MLE for $\pi$ is $\bar{Z}$, and hence the MLE for $p$ is:

$$\hat{p} = 1 - \left( \frac{\bar{Z} - S_e}{r} \right)^{1/c}$$

$\hat{p}$ is asymptotically unbiased and asymptotically normal.

# Estimation and Optimality

## Theorem (Maximum Likelihood Estimator)

Let $r = S_p + S_e - 1$ and $\pi = \Pr[Z_j = 1] = S_e + r(1-p)^c$. Then the MLE for $\pi$ is $\bar{Z}$, and hence the MLE for $p$ is:

$$\hat{p} = 1 - \left(\frac{\bar{Z} - S_e}{r}\right)^{1/c}$$

$\hat{p}$ is asymptotically unbiased and asymptotically normal.

## Theorem (Asymptotic Variance of MLE)

Considering $J = N/c$ fixed, $\hat{p}$ has asymptotic variance:

$$\sigma^2(p, c, J) = \frac{(S_e - r(1-p)^c)\left(r(1-p)^c + 1 - S_e\right)}{Jc^2r^2(1-p)^{2(c-1)}}$$

# Estimation and Optimality

**Theorem (Optimality of Pooled Samples) [Liu et al., 2012]**

Holding $J = N/c$ fixed, for any $c$, there is an upper bound $p_{\max}$ such that $\sigma^2(p, c, J) \leq \sigma^2(p, 1, J)$ when $p \leq p_{\max}$.

# Estimation and Optimality

**Theorem (Optimality of Pooled Samples) [Liu et al., 2012]**

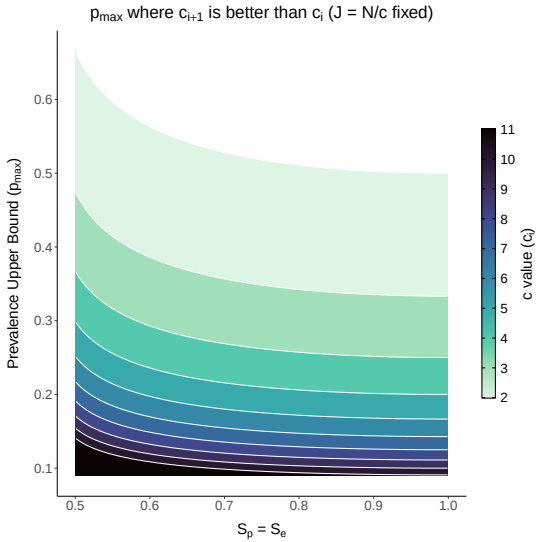
Holding $J = N/c$ fixed, for any $c$, there is an upper bound $p_{\max}$ such that $\sigma^2(p, c, J) \leq \sigma^2(p, 1, J)$ when $p \leq p_{\max}$.

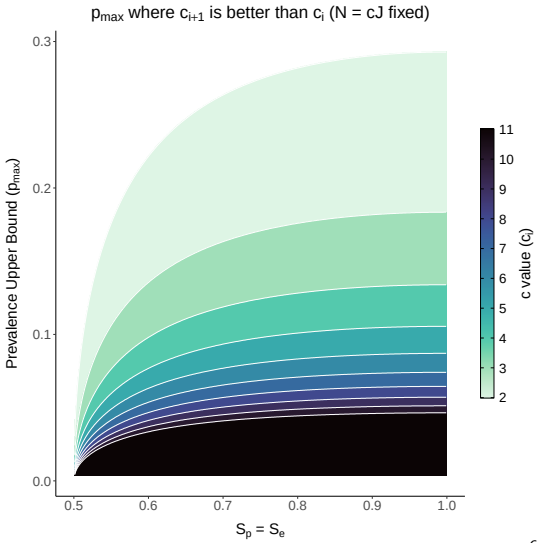
Identical results can be shown when $N$ is fixed.

**Theorem (Optimality of Pooled Samples) [Liu et al., 2012]**

Holding $N = cJ$ fixed, for any $c$, there is an upper bound $p_{\max}$ such that $\sigma^2(p, c, N) \leq \sigma^2(p, 1, N)$ when $p \leq p_{\max}$.

# Estimation and Optimality



# $\varepsilon$-Differential Privacy

**Definition ($\varepsilon$-Differential Privacy) [Dwork and Roth, 2014]**

A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is $\varepsilon$-**differentially private** if, for all $X, X' \in \mathcal{X}$ that differ by 1 (under some metric) and all $Y \in \mathcal{Y}$,

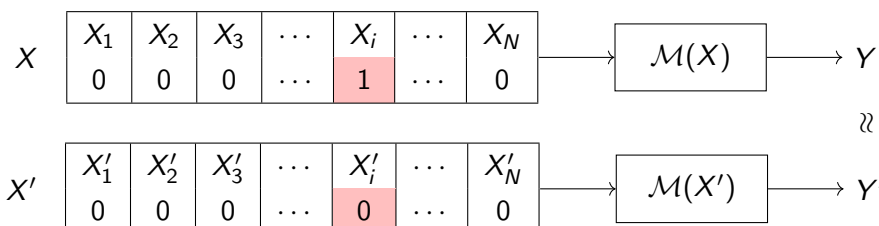
$$\frac{\Pr[\mathcal{M}(X) = Y]}{\Pr[\mathcal{M}(X') = Y]} \leq \exp(\varepsilon)$$

# $\varepsilon$-Differential Privacy

## Definition ($\varepsilon$-Differential Privacy) [Dwork and Roth, 2014]

A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is $\varepsilon$-**differentially private** if, for all $X, X' \in \mathcal{X}$ that differ by 1 (under some metric) and all $Y \in \mathcal{Y}$,

$$\frac{\Pr[\mathcal{M}(X) = Y]}{\Pr[\mathcal{M}(X') = Y]} \leq \exp(\varepsilon)$$



# $\varepsilon$-Differential Privacy

**Definition (Randomized Response Mechanism)**

The **Randomized Response** (RR) mechanism $\mathcal{M}_{\text{RR}} : \{0, 1\} \to \{0, 1\}$ is defined as:

$$\mathcal{M}_{\text{RR}}(X) = \begin{cases} X, & \text{w.p. } \Gamma \\ 1 - X, & \text{w.p. } 1 - \Gamma \end{cases} \quad \text{where } \Gamma \in (0.5, 1)$$

# $\varepsilon$-Differential Privacy

## Definition (Randomized Response Mechanism)

The **Randomized Response** (RR) mechanism $\mathcal{M}_{\mathsf{RR}} : \{0,1\} \to \{0,1\}$ is defined as:

$$\mathcal{M}_{\mathsf{RR}}(X) = \begin{cases} X, & \text{w.p. } \Gamma \\ 1 - X, & \text{w.p. } 1 - \Gamma \end{cases} \quad \text{where } \Gamma \in (0.5, 1)$$

## Theorem (DP of RR)

RR is $\log\left(\frac{\Gamma}{1-\Gamma}\right)$-differentially private.

$$\max\left\{\frac{\Pr[\mathcal{M}_{\mathsf{RR}}(0)=0]}{\Pr[\mathcal{M}_{\mathsf{RR}}(1)=0]}, \frac{\Pr[\mathcal{M}_{\mathsf{RR}}(0)=1]}{\Pr[\mathcal{M}_{\mathsf{RR}}(1)=1]}\right\} = \max\left\{\frac{\Gamma}{1-\Gamma}, \frac{1-\Gamma}{\Gamma}\right\} \leq \exp(\varepsilon)$$

# $\varepsilon$-Differential Privacy

Error in a test is equivalent to two RR mechanisms, defined by sensitivity $S_e$ and specificity $S_p$.

$$\text{True State} \begin{cases} X_i = 1 \xrightarrow{\text{Diagnostic Test}} Y_i = \begin{cases} 1, & \text{w.p. } S_e \\ 0, & \text{w.p. } 1 - S_e \end{cases} \\ X_i = 0 \xrightarrow{\text{Diagnostic Test}} Y_i = \begin{cases} 0, & \text{w.p. } S_p \\ 1, & \text{w.p. } 1 - S_p \end{cases} \end{cases} \Bigg\} \text{Observed State}$$

# $\varepsilon$-Differential Privacy

Error in a test is equivalent to two RR mechanisms, defined by sensitivity $S_e$ and specificity $S_p$.

$$\text{True State} \begin{cases} X_i = 1 \xrightarrow{\text{Diagnostic Test}} Y_i = \begin{cases} 1, & \text{w.p. } S_e \\ 0, & \text{w.p. } 1 - S_e \end{cases} \\ X_i = 0 \xrightarrow{\text{Diagnostic Test}} Y_i = \begin{cases} 0, & \text{w.p. } S_p \\ 1, & \text{w.p. } 1 - S_p \end{cases} \end{cases} \Bigg\} \text{Observed State}$$

**Theorem (DP of Unpooled Samples)**

Testing with sensitivity $S_e$ and specificity $S_p$ satisfies $\varepsilon$-differential privacy, where:

$$\exp(\varepsilon) = \begin{cases} \frac{S_e}{1-S_p}, & S_p > S_e \\ \frac{S_p}{1-S_e}, & S_p \leq S_e \end{cases}$$

# Differential Privacy of Sample Pooling

## Theorem (DP of Sample Pooling)

Sample pooling with misclassification satisfies $\varepsilon$-differential privacy, with:

$$\exp(\varepsilon) = \begin{cases} \gamma(p) & \text{if } S_p > S_e \\ \alpha & \text{if } S_p \leq S_e \end{cases} \quad \text{where} \quad \gamma(p) = \begin{cases} 1/\beta, & p \leq p_{\text{crit},1} \\ \alpha, & p_{\text{crit},1} < p < p_{\text{crit},2} \\ 1/\alpha, & p \geq p_{\text{crit},2} \end{cases}$$

$$\alpha = \frac{1 - S_p + (1-p)^{c-1}(2S_p - 1)}{1 - S_e}, \quad \beta = \frac{S_p - (1-p)^{c-1}(2S_p - 1)}{S_e}$$

# Differential Privacy of Sample Pooling

## Theorem (DP of Sample Pooling)

Sample pooling with misclassification satisfies $\varepsilon$-differential privacy, with:

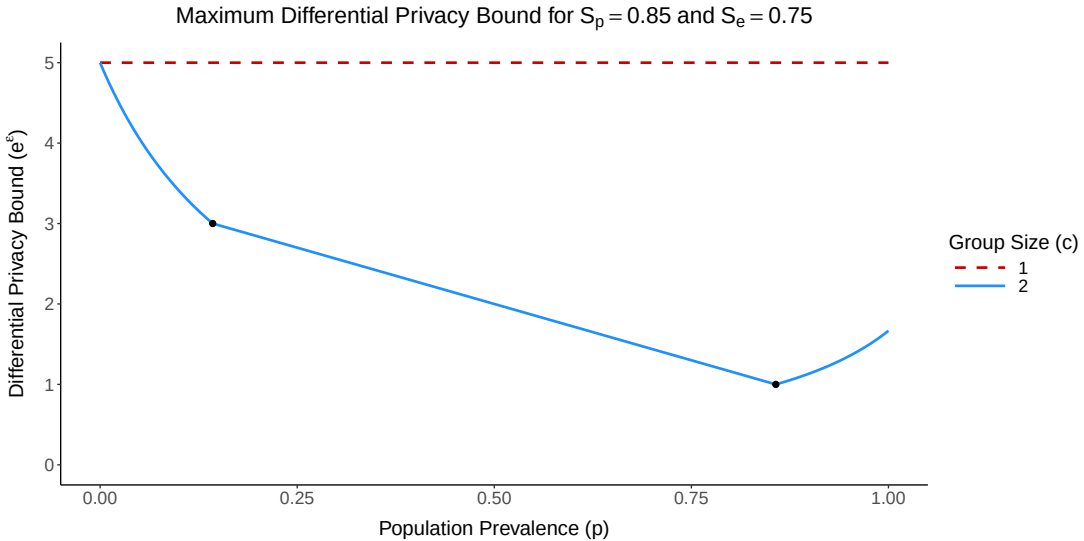
$$\exp(\varepsilon) = \begin{cases} \gamma(p) & \text{if } S_p > S_e \\ \alpha & \text{if } S_p \leq S_e \end{cases} \quad \text{where} \quad \gamma(p) = \begin{cases} 1/\beta, & p \leq p_{\text{crit},1} \\ \alpha, & p_{\text{crit},1} < p < p_{\text{crit},2} \\ 1/\alpha, & p \geq p_{\text{crit},2} \end{cases}$$

$$\alpha = \frac{1 - S_p + (1-p)^{c-1}(2S_p - 1)}{1 - S_e}, \quad \beta = \frac{S_p - (1-p)^{c-1}(2S_p - 1)}{S_e}$$

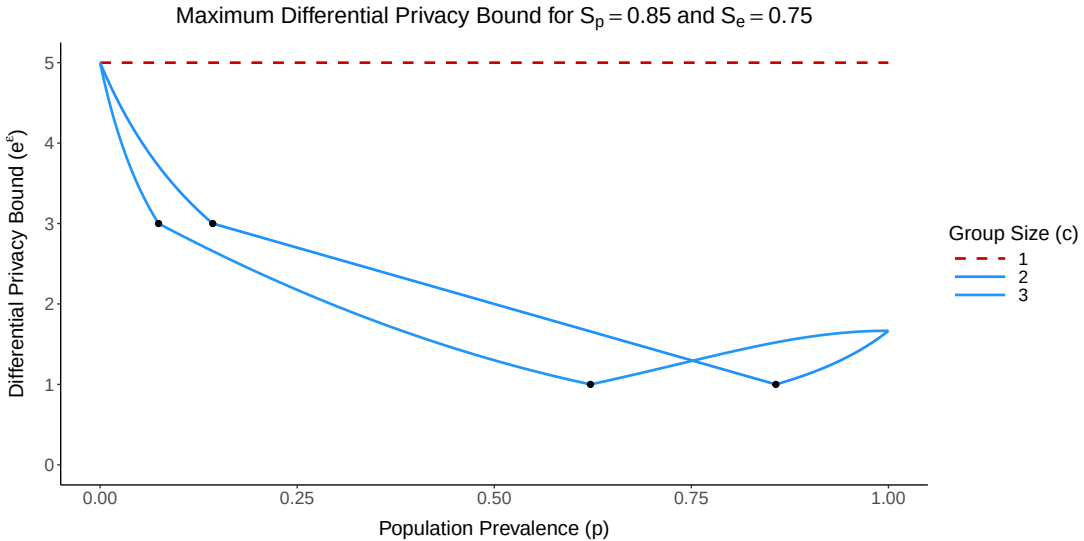
## Theorem (DP Optimality)

For all $p > 0$, pooled samples provide strictly better differential privacy than unpooled samples. Also, if $S_p > S_e$, $p = p_{\text{crit},2}$ results in the lowest $\varepsilon$.

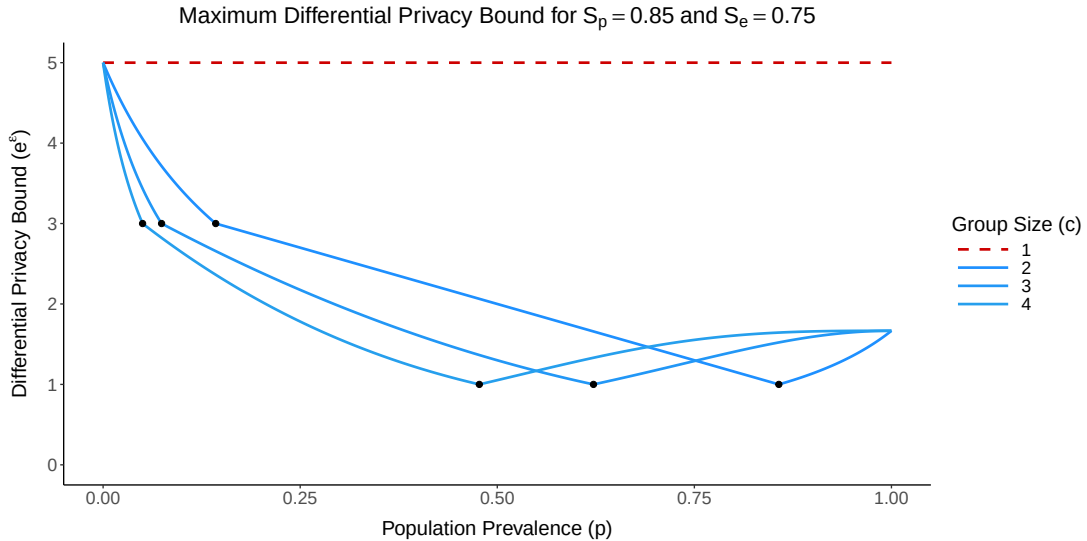
# Differential Privacy of Sample Pooling



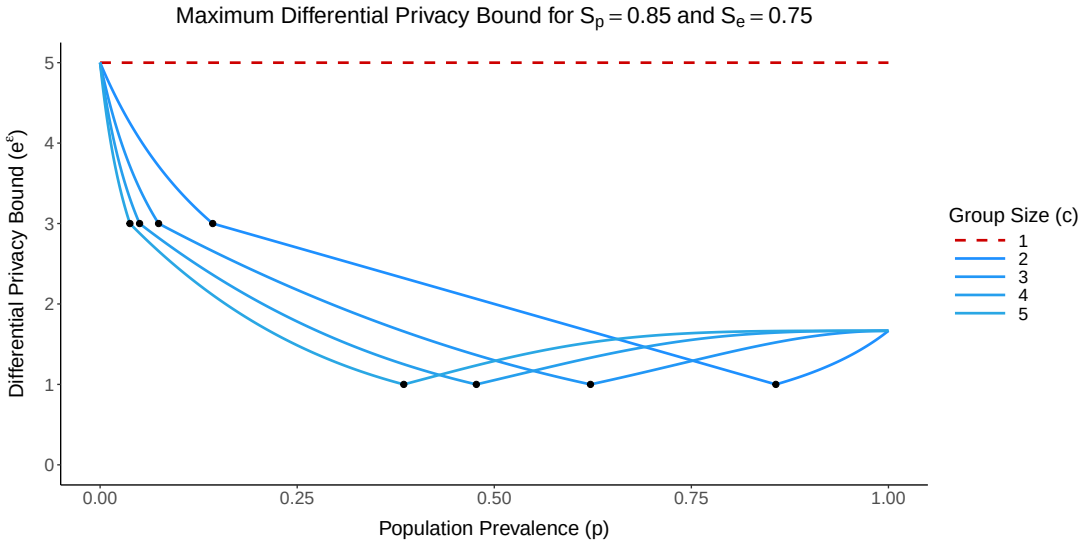
# Differential Privacy of Sample Pooling



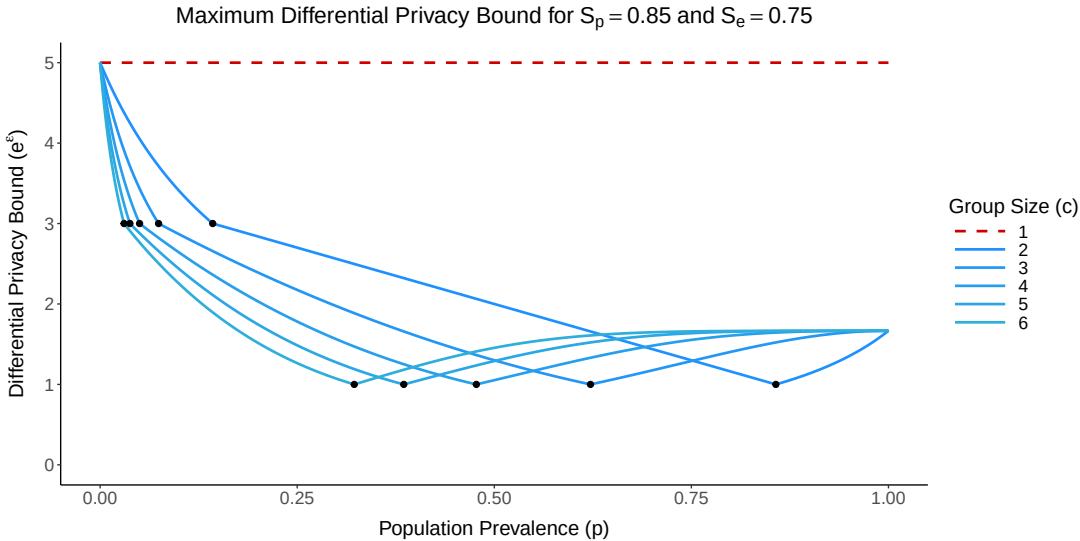
# Differential Privacy of Sample Pooling



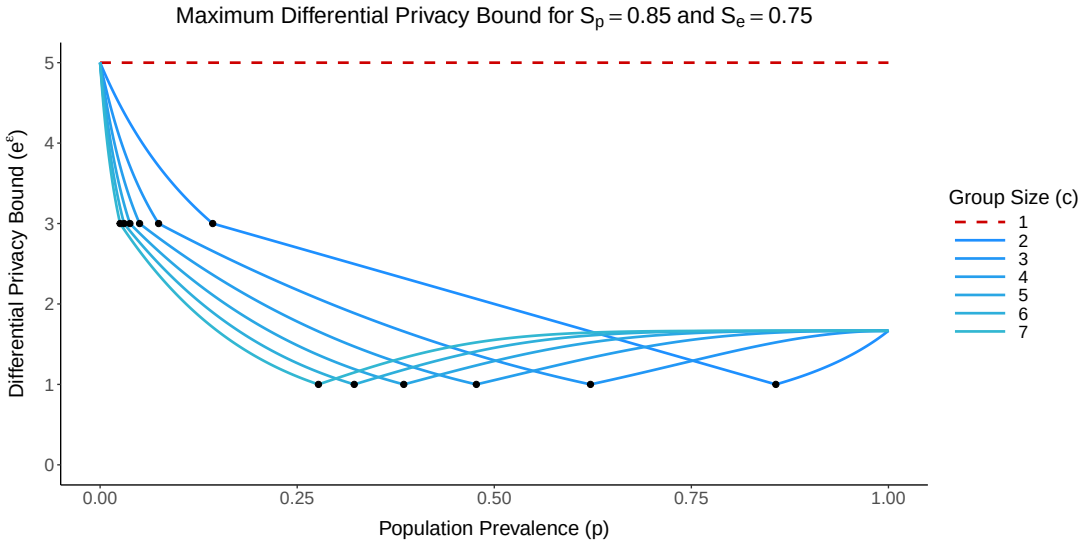
# Differential Privacy of Sample Pooling



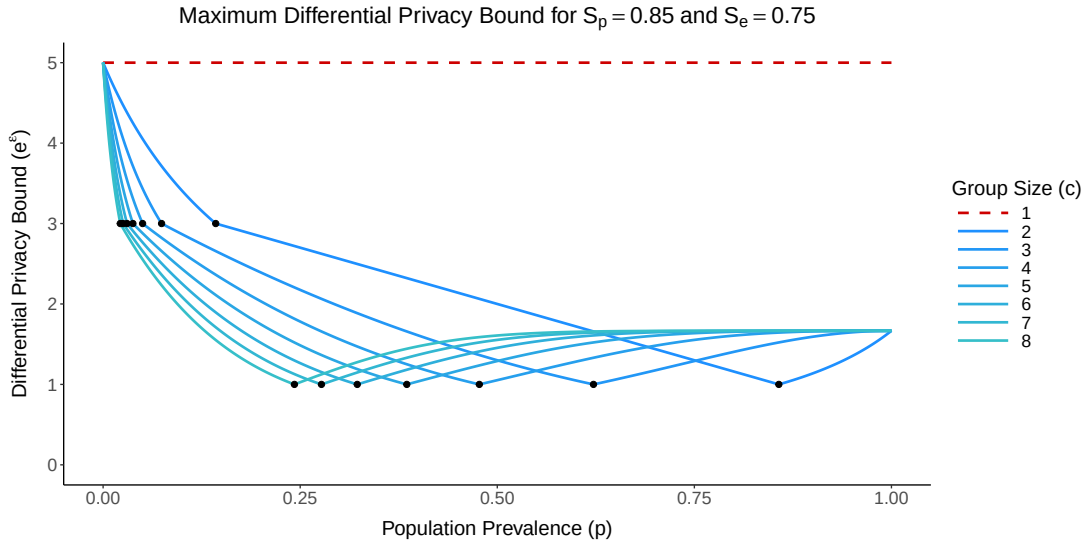
# Differential Privacy of Sample Pooling



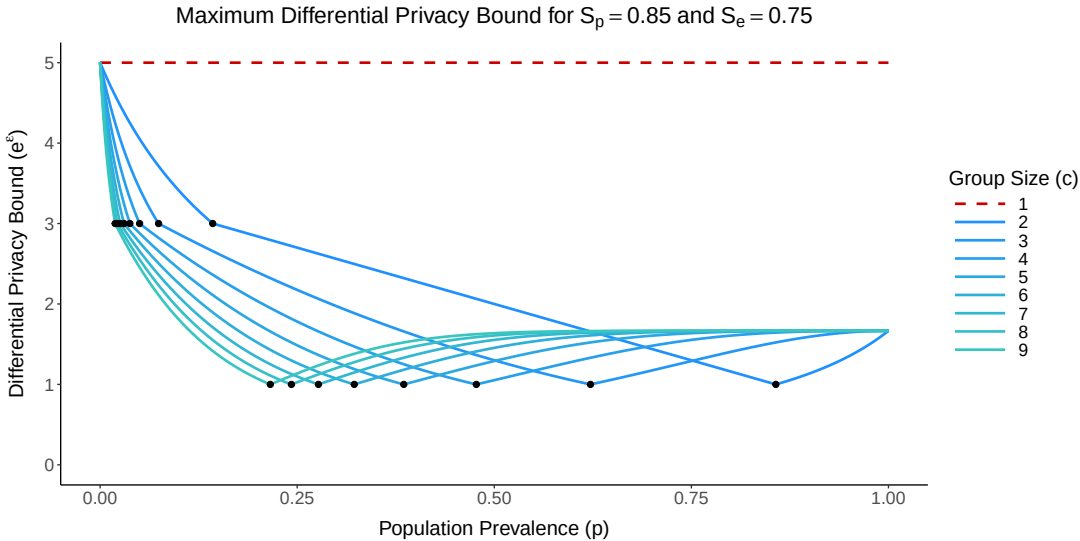
# Differential Privacy of Sample Pooling



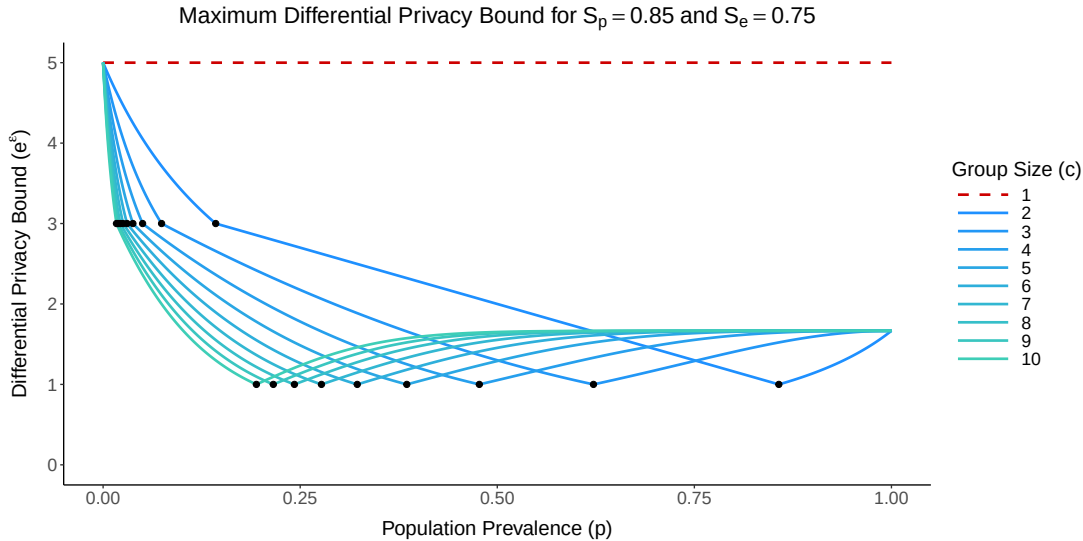
# Differential Privacy of Sample Pooling



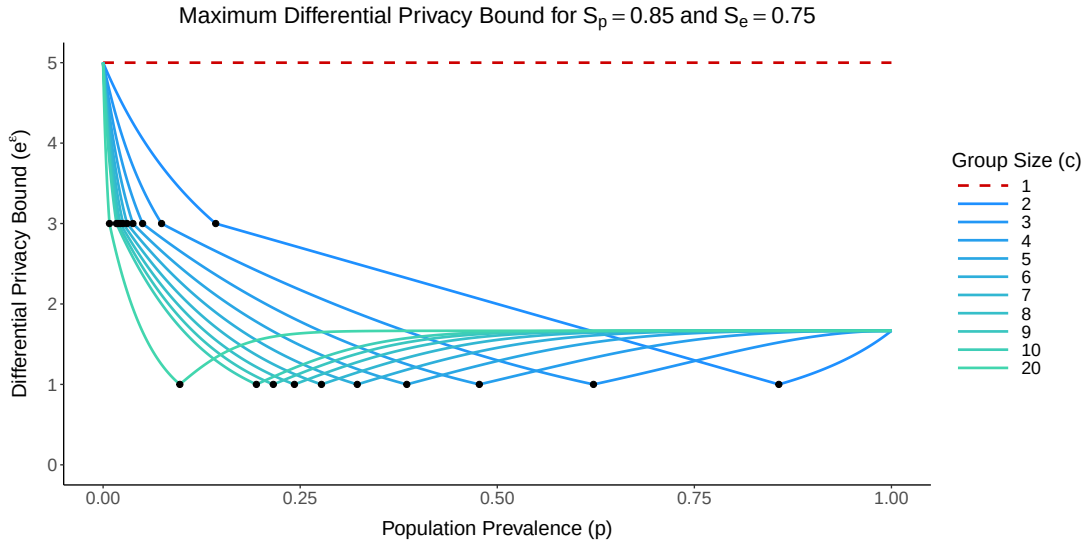
# Differential Privacy of Sample Pooling



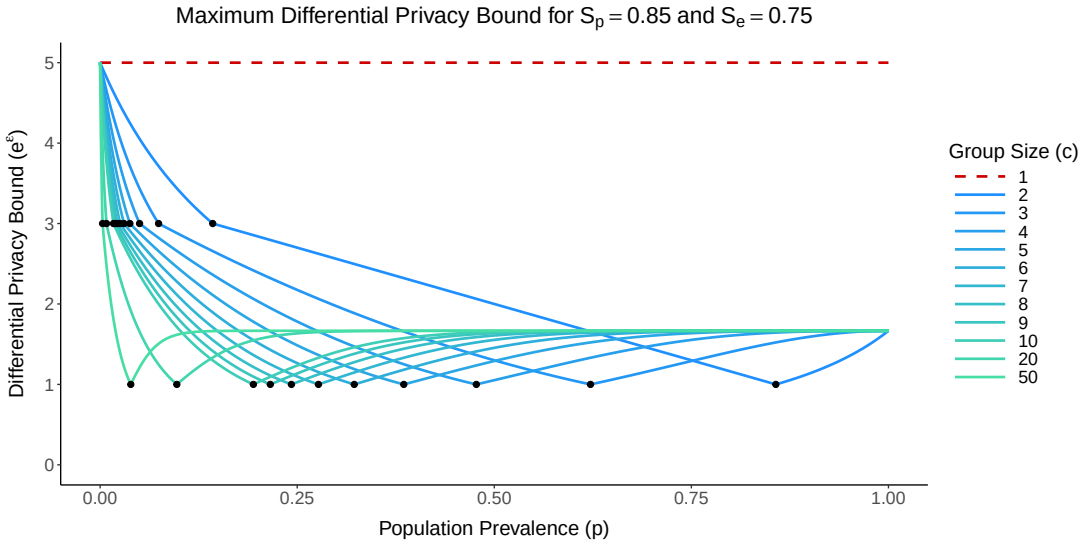
# Differential Privacy of Sample Pooling



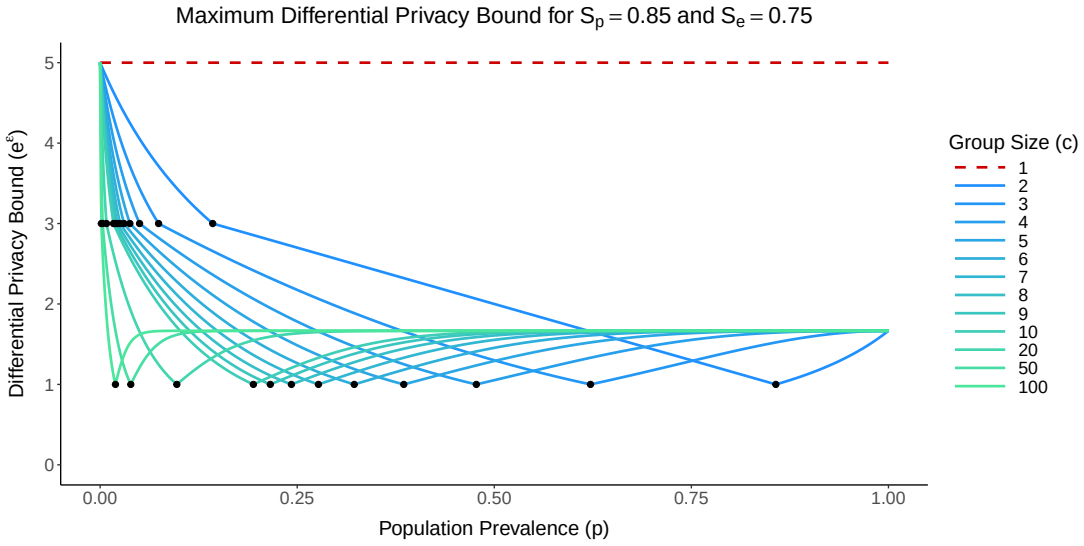
# Differential Privacy of Sample Pooling



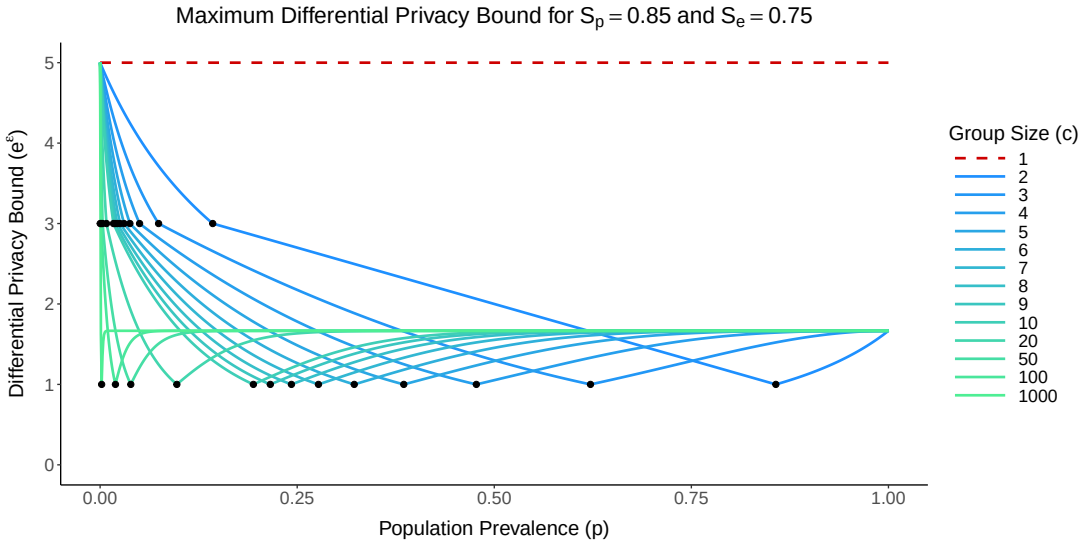
# Differential Privacy of Sample Pooling



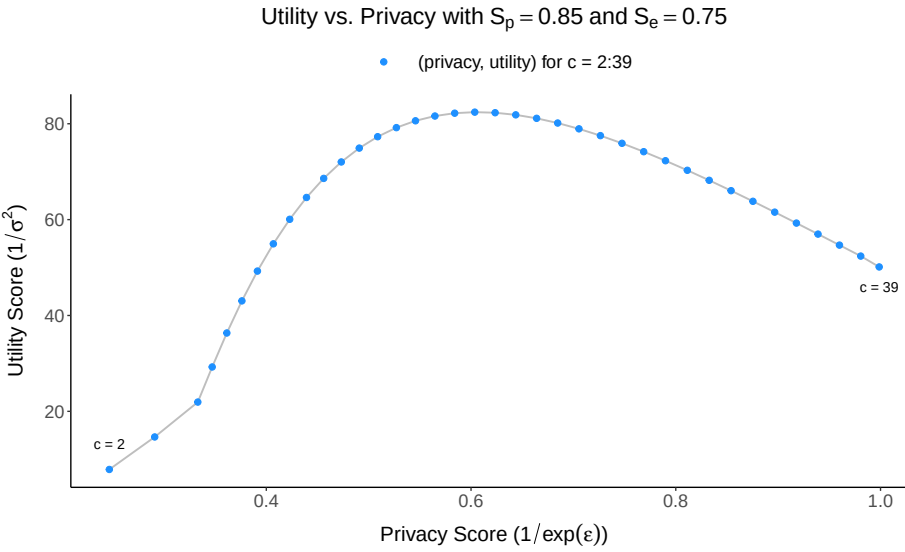
# Differential Privacy of Sample Pooling



# Differential Privacy of Sample Pooling



# Privacy and Estimation



# Summary

With a reasonable upper bound on $p$, sample pooling means:

- $J = N/c$ tests required, instead of $N$
- Far better differential privacy, which is a function of $p$ and $c$, not just $S_p$ and $S_e$
- Variance of pooled estimator is lower than variance of unpooled estimator

# Summary

With a reasonable upper bound on $p$, sample pooling means:

- $J = N/c$ tests required, instead of $N$
- Far better differential privacy, which is a function of $p$ and $c$, not just $S_p$ and $S_e$
- Variance of pooled estimator is lower than variance of unpooled estimator

If $S_e$ and $S_p$ are fixed, we can artificially decrease them via RR.

# References


Dwork, C. and Roth, A. (2014).
The Algorithmic Foundations of Differential Privacy.
*Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Liu, A., Liu, C., Zhang, Z., and Albert, P. S. (2012).
Optimality of group testing in the presence of misclassification.
*Biometrika*, 99(1):245–251.