Differentially Private Chi-Squared Hypothesis Testing

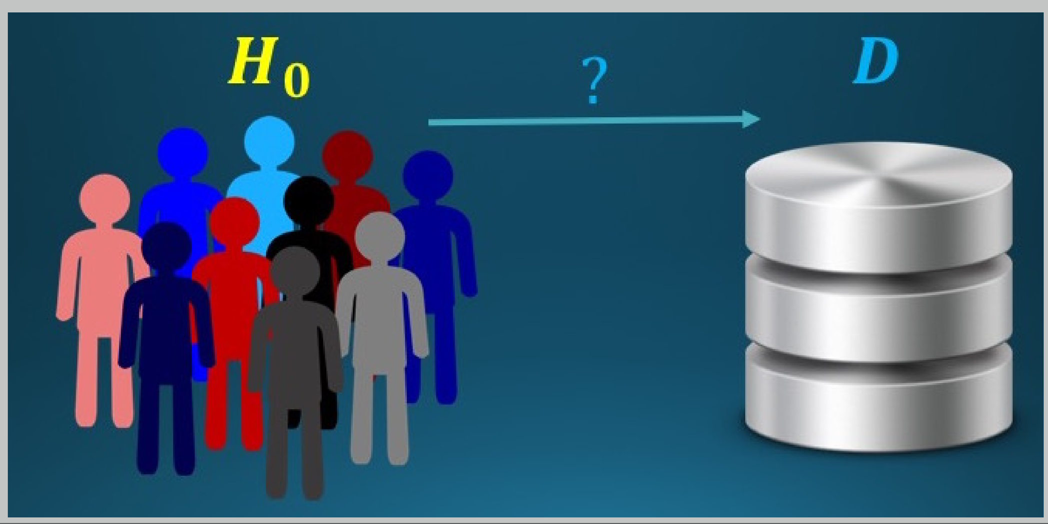
Privacy Tools for Sharing Research Data
A National Science Foundation Secure and Trustworthy Cyberspace Project

Marco Gaboardi, Hyun woo Lim, **Ryan Rogers**, and Salil Vadhan

Hypothesis Testing

- Given dataset D and proposed model of the data H_0, we want to determine whether H_0 should be rejected or not.
- **Goal**: Design a test that leads to small Type I and Type II error.

	H_0 True	H_0 False
Reject H_0	α Type I	$1-\beta$
Not	$1-\alpha$	β Type II



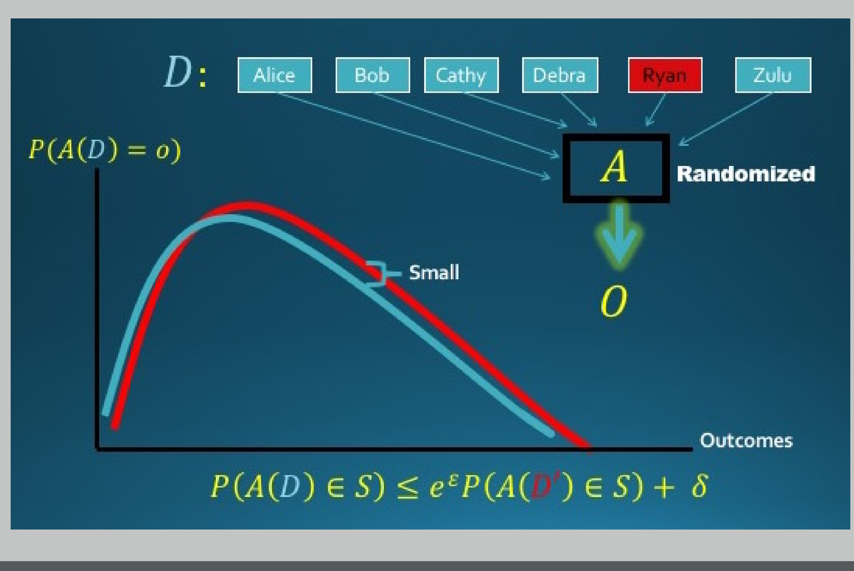
The Need for Privacy

- Data may contain sensitive information, e.g. medical data
- Releasing the result may leak information
- Homer et al. '08 showed that with only aggregate statistics on *genomic-wide association studies* can determine whether someone in the study has a disease or not.

New Goal: Obtain statistically valid hypothesis tests which preserve the privacy of those in the study.

Differential Privacy

Outcome of test $A : \mathcal{D} \to \mathcal{O}$ should *roughly* stay the same if one person changes his data.



Focus of this work: Chi-Square Tests

- Categorical data $X \sim \text{Multinomial}(n, \mathbf{p})$ where $\mathbf{p} = (p_1, \cdots, p_d)$
- Tests using the chi-square statistic:

$$Q^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}.$$

- **Goodness of Fit**: $H_0 : \mathbf{p} = \mathbf{p}^0$.
- **Independence Testing**: $Y^1 \sim \text{Multinomial}(1, \pi^1)$ and $Y^2 \sim \text{Multinomial}(1, \pi^2)$ are independent. Form the contingency table of counts based on n trials:

	$Y^2 = 0$	$Y^2 = 1$
$Y^1 = 0$	X_{00}	X_{01}
$Y^1 = 1$	X_{10}	X_{11}

- Tests based on a *critical value* τ, so that if $Q^2 > \tau$ then Reject H_0.
- Known that $Q^2 \xrightarrow{D} \chi^2_{df}$, so we set $\tau = \chi^2_{df,1-\alpha}$ in order for Type I error to be nearly α. Works well even for moderately sized datasets.

Prior Work for DP Hypothesis Tests

- Add independent noise Z to each cell count and use the classical test with $Q^2_{DP} = Q^2(X + Z)$.
- Scale of noise is small, but how does it perform?
- Set $\alpha = 5\%$, $\epsilon = 0.1$.

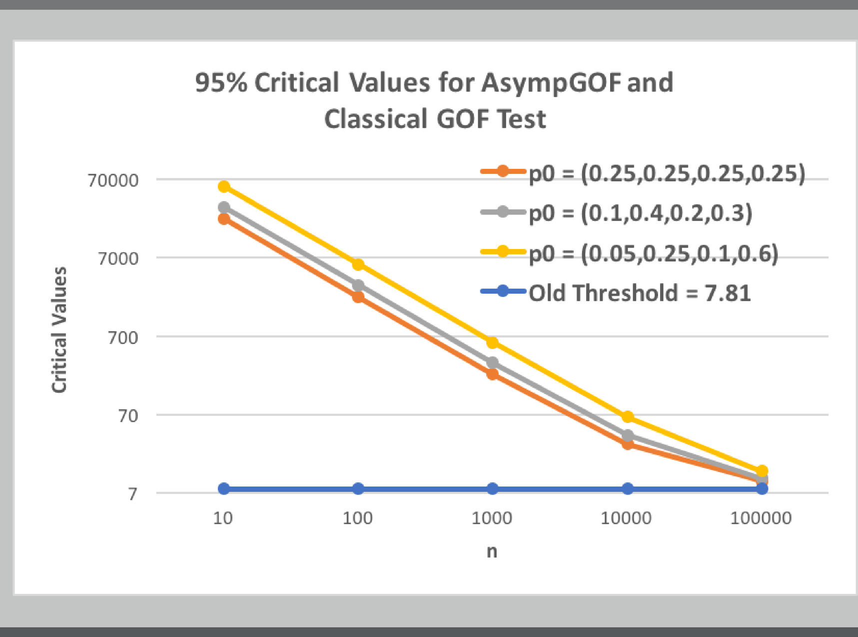
$\vec{p}^0$	n	$\chi^2_{df,1-\alpha}$	Type I error
(.25,.25,.25,.25)	100	7.81	100%
(.25,.25,.25,.25)	1,000	7.81	99%
(.25,.25,.25,.25)	10,000	7.81	65%
(.25,.25,.25,.25)	100,000	7.81	10%
(.1,.2,.3,.4)	10,000	7.81	70%
(.1,.2,.3,.4)	100,000	7.81	12%

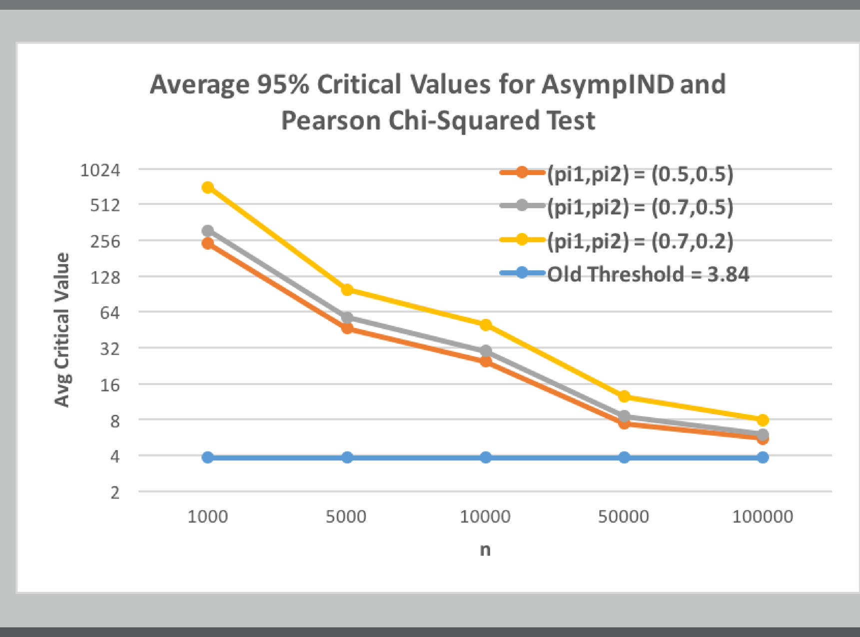
Our Contribution for DP Hypothesis Tests

- Add noise to cell counts to get Q^2_{DP}, incorporate the distribution from the additional noise when computing a new critical value.
- New GOF and independence tests:
  - MC based tests which can use either Laplace or Gaussian noise. GOF test provably achieves at most target Type I error.
  - Tests based on new asymptotic distribution (`AsympGOF` and `AsympIND`)- only for Gaussian noise.

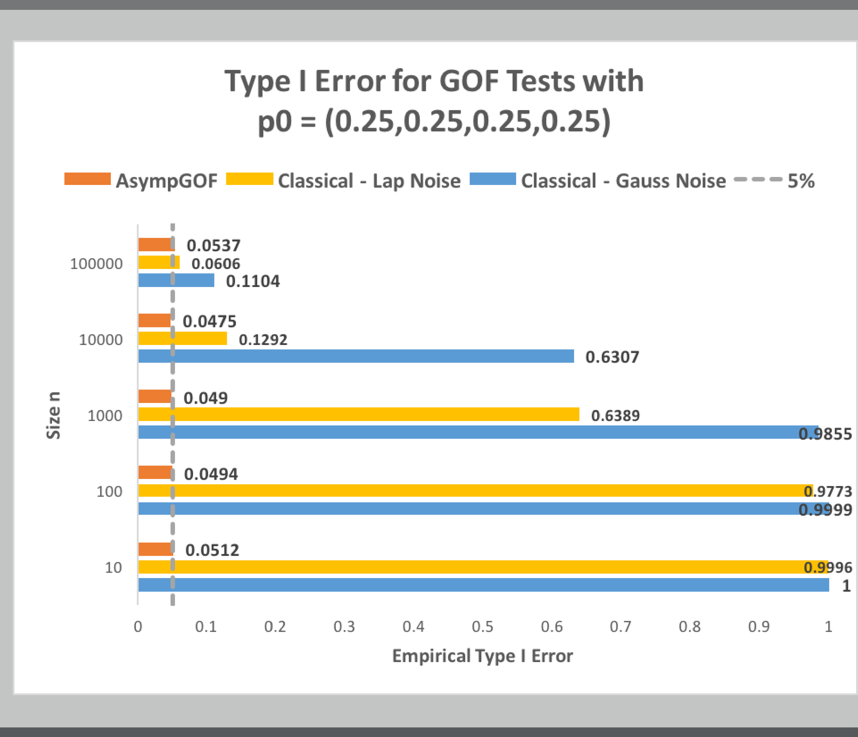
$$Q^2_{DP} \xrightarrow{D} \sum_i \lambda_i \chi^2_1 \quad \text{Use this to compute new critical values.}$$

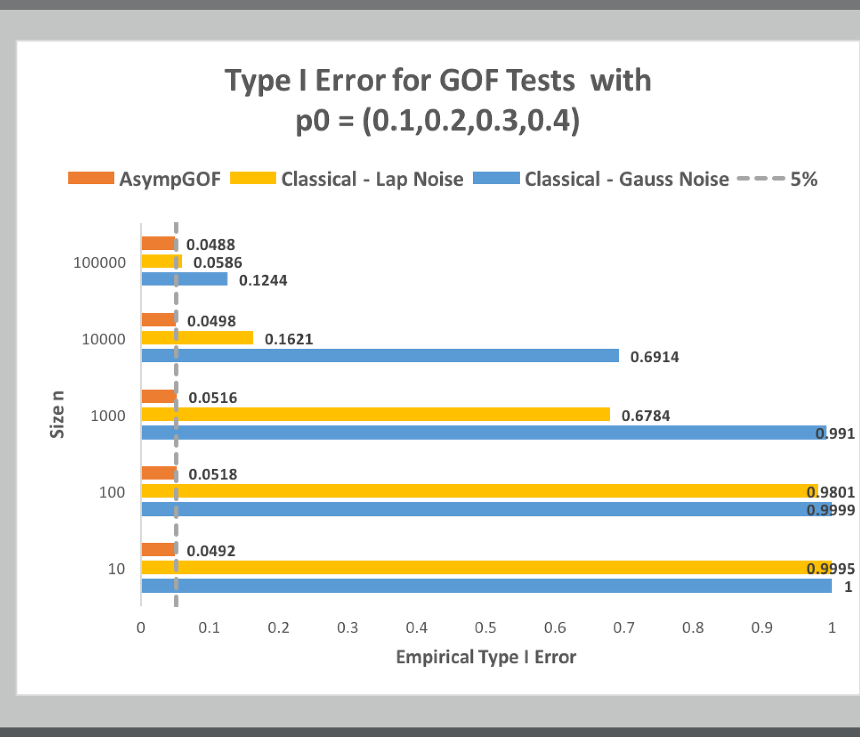
Modified Critical Values for $\alpha = 0.05, \epsilon = 0.1$



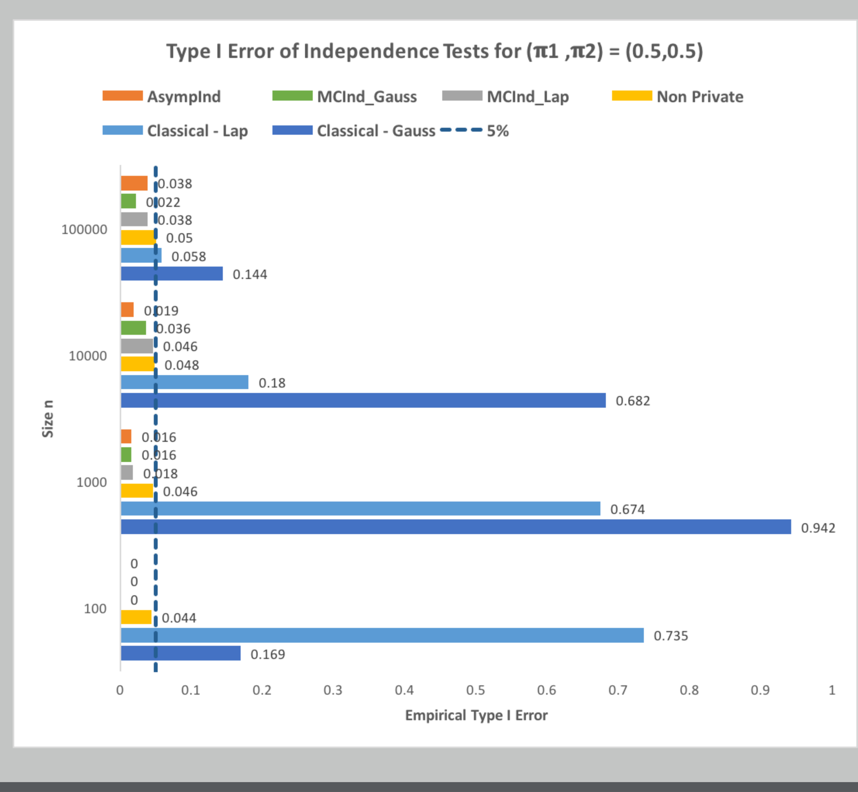


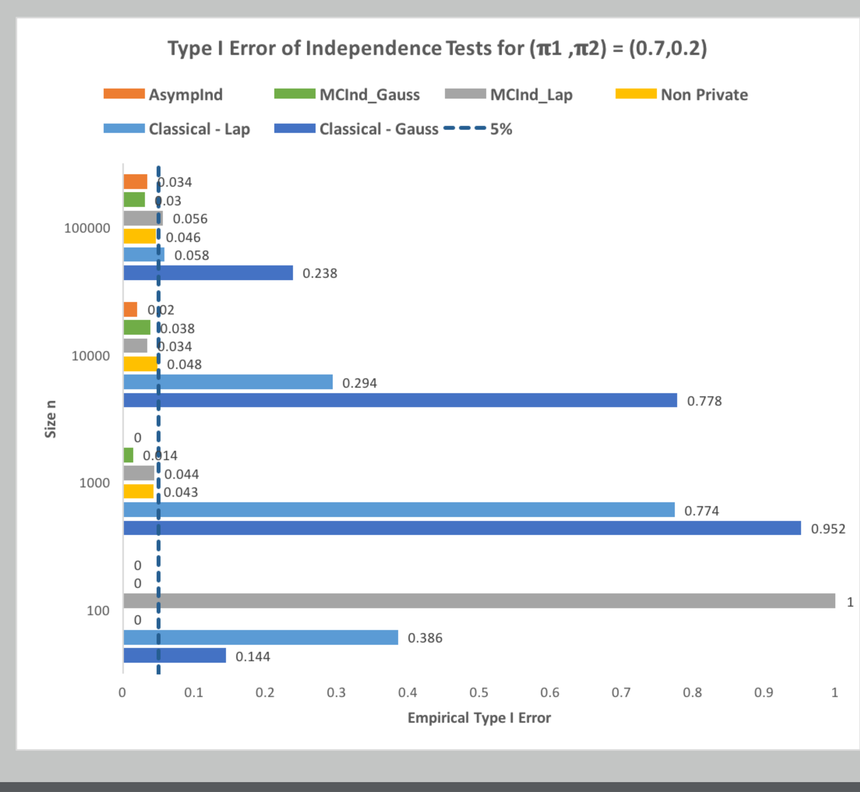
Type I Error for Private GOF Tests: $\alpha = 0.05, \epsilon = 0.1$





Type I Error for Private Independence Tests: $\alpha = 0.05, \epsilon = 0.1$





Type II Error: Data not generated from H_0

For GOF we sample X with $\mathbf{p}^0 + (0.01) \cdot (1, -1, 1, -1)$.

For independence we add covariance 0.01 between Y^1 and Y^2.