Differentially Private Chi-Squared Hypothesis Testing:

Goodness of Fit and Independence Testing

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

Classical Hypothesis Testing

Classical Hypothesis Testing

- Given two models of a population H_0 and H_1, which model is best supported by a sample from the population?
- Want to design a test $A: X^n \rightarrow \{Reject\ H_0, Fail\ to\ Reject\ H_0\}$ such that

	H_0 is True	H_1 is True
A rejects	α	1-β
A fails to reject	1-α	β

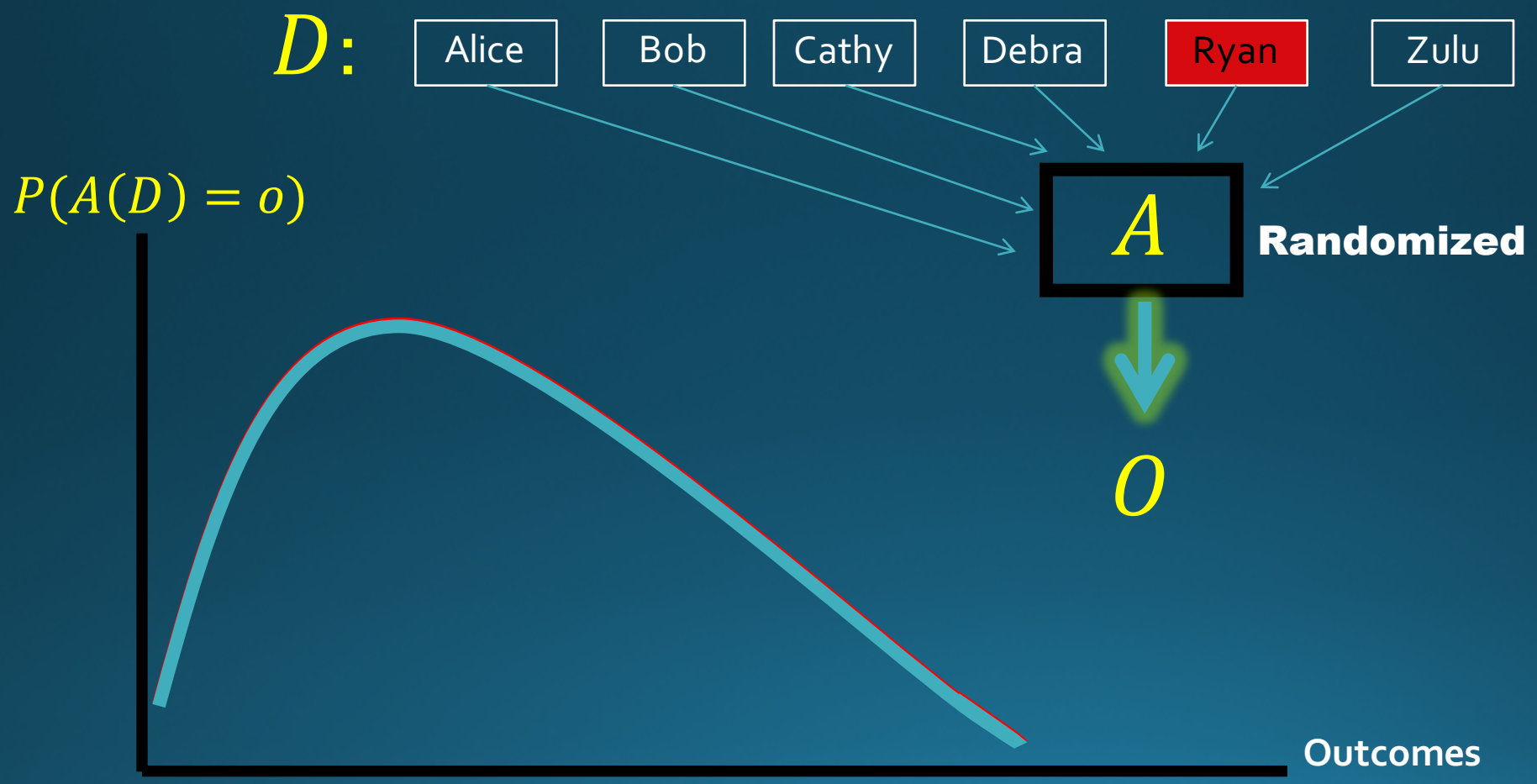
- Want to ensure test has ***significance*** at least $1 - \alpha$ and then hope to maximize ***power***.

Data may be Sensitive!

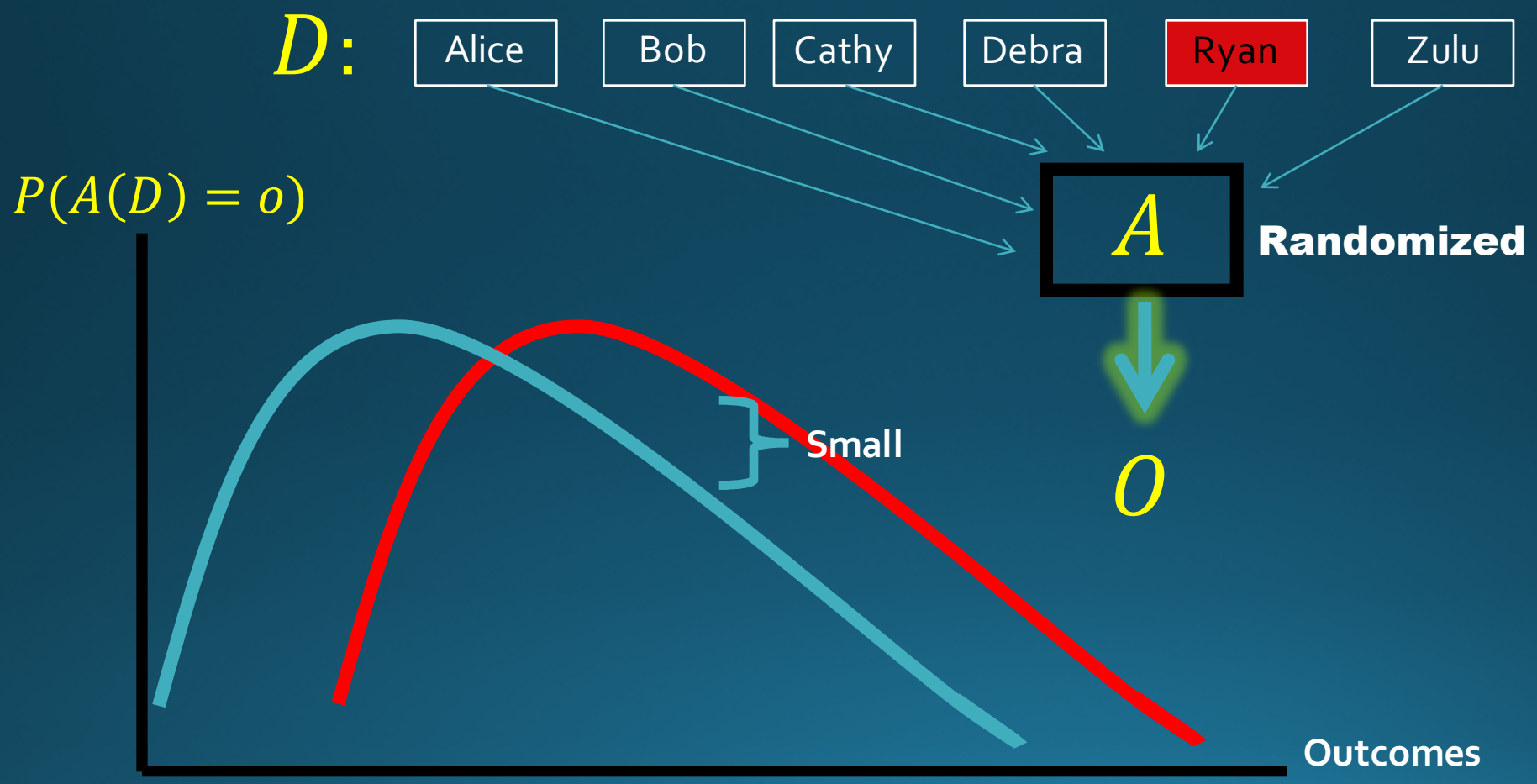
- Data may contain highly sensitive information, e.g. medical information.
- Releasing a result of a hypothesis test may leak information about the individuals in the data.
- Homer et al. 2008 showed that from aggregate statistics on GWAS data, one can detect whether a particular individual was in the dataset.
- Can we still do hypothesis testing while preserving the privacy of the individuals?

Differential Privacy [DMNS '06]

Differential Privacy [DMNS '06]



Differential Privacy [DMNS '06]



Differential Privacy [DMNS '06]

- A randomized algorithm $A: X^n \to Y$ is (ε, δ)-differentially private if for any neighboring data sets $D, D' \in X^n$ and for any outcome $S \subseteq Y$ we have

$$P(A(D) \in S) \leq e^{\varepsilon} P(A(D) \in S) + \delta$$

Focus of this Work

- Categorical data: $D = (D_1, \ldots, D_d) \sim Multinomial(n, \vec{p})$.

1. **Goodness of Fit**: $H_0: \vec{p} = \vec{p}^0$
   - Simple Test - data distribution completely determined
2. **Independence Test**: $H_0: Y^{(1)} \perp Y^{(2)}$
   - Composite Test – data distribution not completely determined

	$Y^{(2)} = 1$	$Y^{(2)} = 0$
$Y^{(1)} = 1$	D_{11}	D_{10}
$Y^{(1)} = 0$	D_{01}	D_{00}

- Both classical tests use the Chi-Squared Statistic:

$$Q^2 = \sum \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Related Work

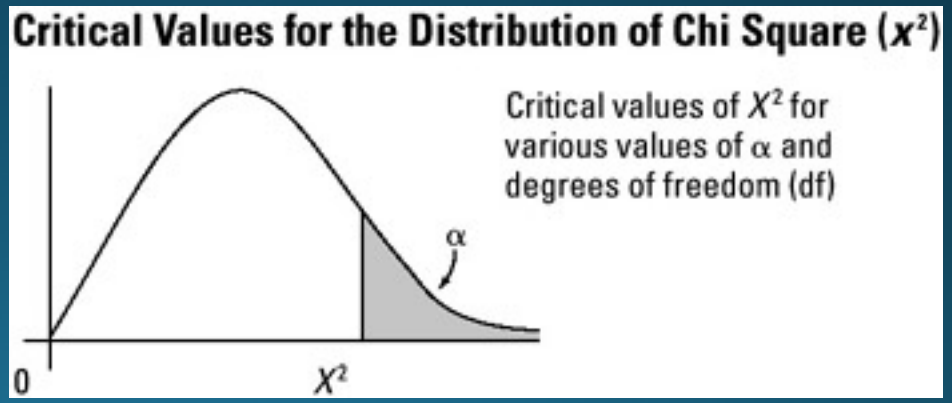
- Smith '11 – general asymptotic result.
- GWAS:
  - Uhler, Slavkovic, and Fienberg '13
  - Yu, Fienberg, Slavkovic, and Uhler '14
  - Johnson and Shmatikov '13
- Using classical tests with noisy data:
  - Vu and Slavkovic '09
  - Fienberg, Rinaldo, Yang '10
  - Karwa and Slavkovic '12, '16
- DP Contingency Tables
  - Barak, Chaudhuri, Dwork, Kale, McSherry, and Talwar '07
  - Li, Hay, Rastogi, Miklau, and McGregor, '10
  - Hardt, Ligett, and McSherry '12
  - Li and Milau '12
  - Gaboardi, Gallego, Hsu, Roth, Wu '14
- Wang, Lee, Kifer '16 – independently also consider GOF and Independence DP Tests

Goodness of Fit Test - Classical

- Null Hypothesis: $H_0 : \vec{p} = \vec{p}^0$
- Form the Chi-squared statistic:

$$Q^2 = \sum_{i=1}^{d} \frac{\left(D_i - np_i^0\right)^2}{np_i^0}$$

- Under H_0, we have $Q^2 \xrightarrow{D} \chi^2_{d-1}$
- Classical Test:
  - If $Q^2 > \chi^2_{d-1,1-\alpha}$ then reject
  - Else, fail to reject.



Private Goodness of Fit – 1st Attempt

- Add noise directly to statistic Q^2.
- To preserve ε-DP, need to add noise with scale $\sim \frac{1}{\varepsilon \min_i\{p_i^0\}}$.
- Noise is too high!

Private Goodness of Fit -2nd attempt

- Add noise to each cell count
- To preserve ε-DP, need to add noise with scale $\sim \frac{2}{\varepsilon}$.
- Form the private chi-squared statistic:

$$Q_{DP}^2 = \sum_{i=1}^{d} \frac{\left(D_i + Z_i - np_i^0\right)^2}{np_i^0}$$ where $Z_i \sim N(0, \sigma^2)$ or $Z_i \sim Lap\left(\frac{2}{\varepsilon}\right)$.

$(\varepsilon, \delta) - DP$ $\varepsilon - DP$

- Still have $Q_{DP}^2 \xrightarrow{D} \chi_{d-1}^2$ for fixed ε, δ.
- Just try the classical test

2nd Attempt Results

- We generate 1000 random samples of multinomial data with size *n* and probability vector $\vec{p}^0$ for various values.
- Fix $1 - \alpha = 0.95$ and privacy parameters $(\varepsilon, \delta) = (0.1, 10^{-6})$.
- Proportion of trials that fell below the $\chi^2_{d-1,1-\alpha}$ gives significance

$\vec{p}^0$	n	$\chi^2_{d-1,1-\alpha}$	Significance
(.25,.25,.25,.25)	100	7.81	0
(.25,.25,.25,.25)	1,000	7.81	0.01
(.25,.25,.25,.25)	10,000	7.81	0.357
(.25,.25,.25,.25)	100,000	7.81	0.901
(.1,.4,.2,.3)	10,000	7.81	0.3
(.1,.4,.2,.3)	100,000	7.81	0.875
(.05,.25,.1,.6)	10,000	7.81	0.168
(.05,.25,.1,.6)	100,000	7.81	0.793
(.01,.29,.1,.6)	100,000	7.81	0.586

Why does classical test do so poorly?

- Fix data and only consider randomness due to noise.
- $E_Z[Q_{DP}^2] = \sum_{i=1}^{d} E_Z\left[\frac{\left(D_i+Z_i-np_i^0\right)^2}{np_i^0}\right] \geq \sum_{i=1}^{d} \frac{1}{np_i^0}\left(E_Z\left[D_i + Z_i - np_i^0\right]\right)^2 = Q^2$.
- This will cause us to reject more often, as our simulations showed.

Outline of Rest of Talk

- Private Goodness of Fit Tests:
  - MC based test with Laplace or Gaussian noise
  - Asymptotic based test with Gaussian noise

- Private Independence Test:
  - MC based test with Laplace or Gaussian noise
  - Asymptotic based test with Gaussian noise

MC Private Goodness of Fit

- With the data, form the private chi-squared statistic Q_{DP}^2 with either Laplace or Gaussian noise.
- Since we know the distribution of the noise and the data under H_0, we can sample points from the distribution of the chi-squared statistic.
- Sample *k* points i.i.d. from distribution of Q_{DP}^2 and sort them.
- The critical value based on these *k* samples is the $\lceil (k+1)(1-\alpha) \rceil$ ranked sample. If Q_{DP}^2 > critical value then reject H_0.
- Test is guaranteed significance at least $1-\alpha$.

Asymptotic Approach to GOF

- We want to obtain a better approximation to the distribution of Q_{DP}^2
- Define the random vector $U = (U_1, U_2, \ldots, U_d)$ where

$$U_i = \frac{D_i - np_i^0}{\sqrt{np_i^0}}.$$

- By the CLT we know that $U \xrightarrow{D} N(0, \Sigma)$ where

$$\Sigma = I - \sqrt{p^0}\sqrt{p^0}^T$$

- Note that $Q^2 = U^T U$

Asymptotic Approach to Private GOF

- We then define the random vector $W = (U, V)$, where U is the same as before and V is the vector of rescaled noise terms

$$V_i = \frac{Z_i}{\sigma}$$

- Note that we can rewrite: $Q_{DP}^2 = W^T A\, W$ where

$$A = \begin{bmatrix} I_d & \Lambda \\ \Lambda & \Lambda^2 \end{bmatrix} \text{ where } \Lambda = Diag\left(\frac{\sigma}{\sqrt{np^0}}\right).$$

- Independently of our work Wang, Lee, and Kifer '16 give this asymptotic distribution when $Z_i \sim Lap\left(\frac{2}{\epsilon}\right)$ and $\epsilon = \Theta\left(\frac{1}{\sqrt{n}}\right)$.
  - However, the resulting distribution is NASTY - quadratic form of Normal-Laplace random variables.
  - May have to rely on MC methods to find critical values for this distribution.

Gaussian Noise

- Recall $U \xrightarrow{D} N(0, \Sigma)$ and $V \sim N(0, I_d)$ thus $W = (U, V) \xrightarrow{D} N(0, \Sigma')$ where

$$\Sigma' = \begin{bmatrix} \Sigma & 0 \\ 0 & I_d \end{bmatrix}$$

- Now we have $Q_{DP}^2 = W^T A\, W$ is a quadratic form of normals
- If $\frac{\sigma}{\sqrt{n}} \to constant$ then

$$Q_{DP}^2 \xrightarrow{D} \sum_i \lambda_i \chi_1^2$$

- Where $\{\lambda_i\}$ are the eigenvalues of $B^T A\, B$ where $BB^T = \Sigma'$.

Private Goodness of Fit - Gaussian

- New Test:
  - Given (ϵ, δ) and $H_0: p = p^0$ find* the critical value τ_ϵ^α where

$$P\left[\sum_i \lambda_i \chi_1^2 > \tau_\epsilon^\alpha\right] = \alpha$$

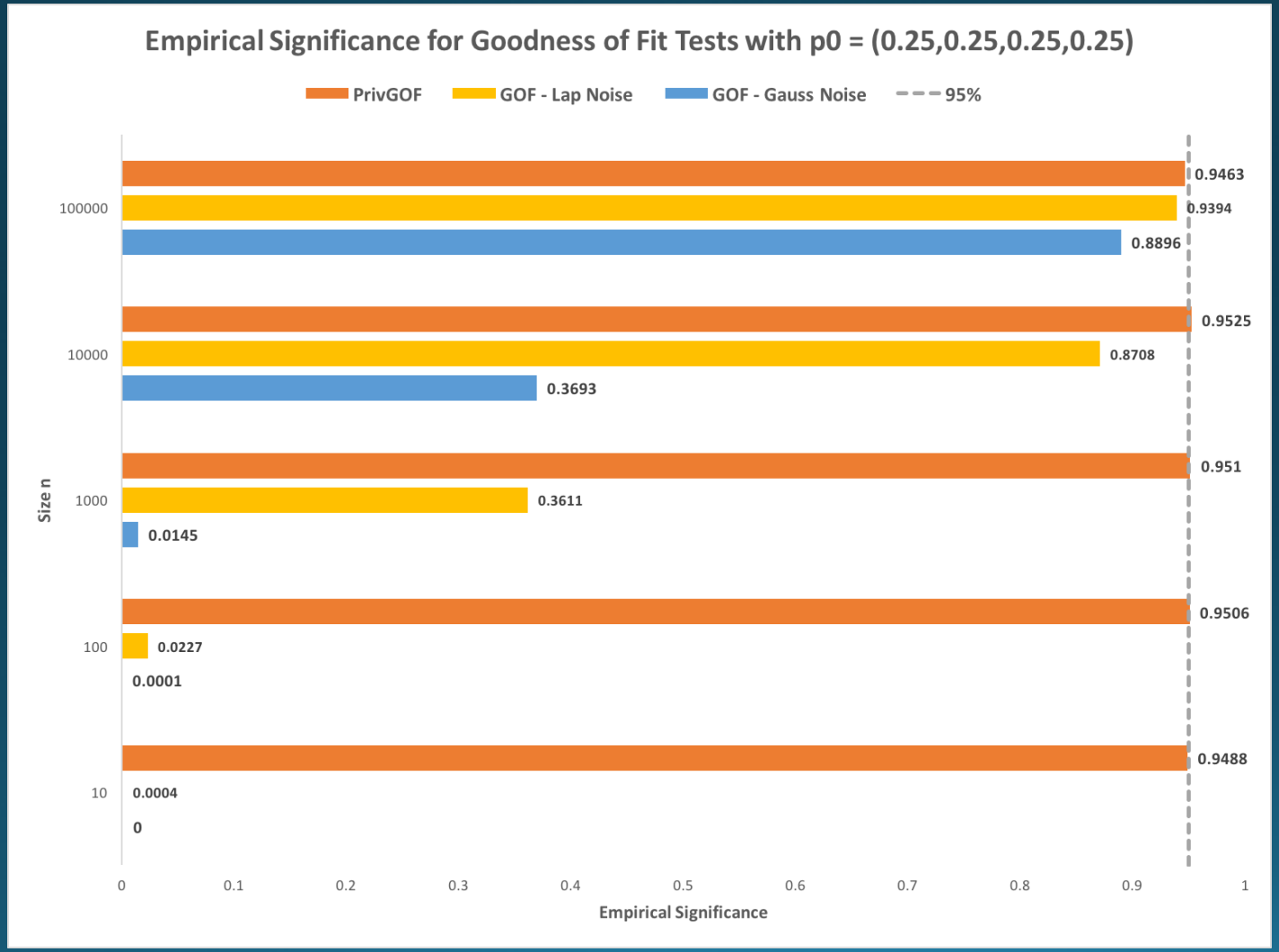
  - If $Q_{DP}^2 > \tau_\epsilon^\alpha$ then reject H_0.
  - Else, fail to reject.

- * We used a numerical solver (CompQuadForm in R) to find the critical values.

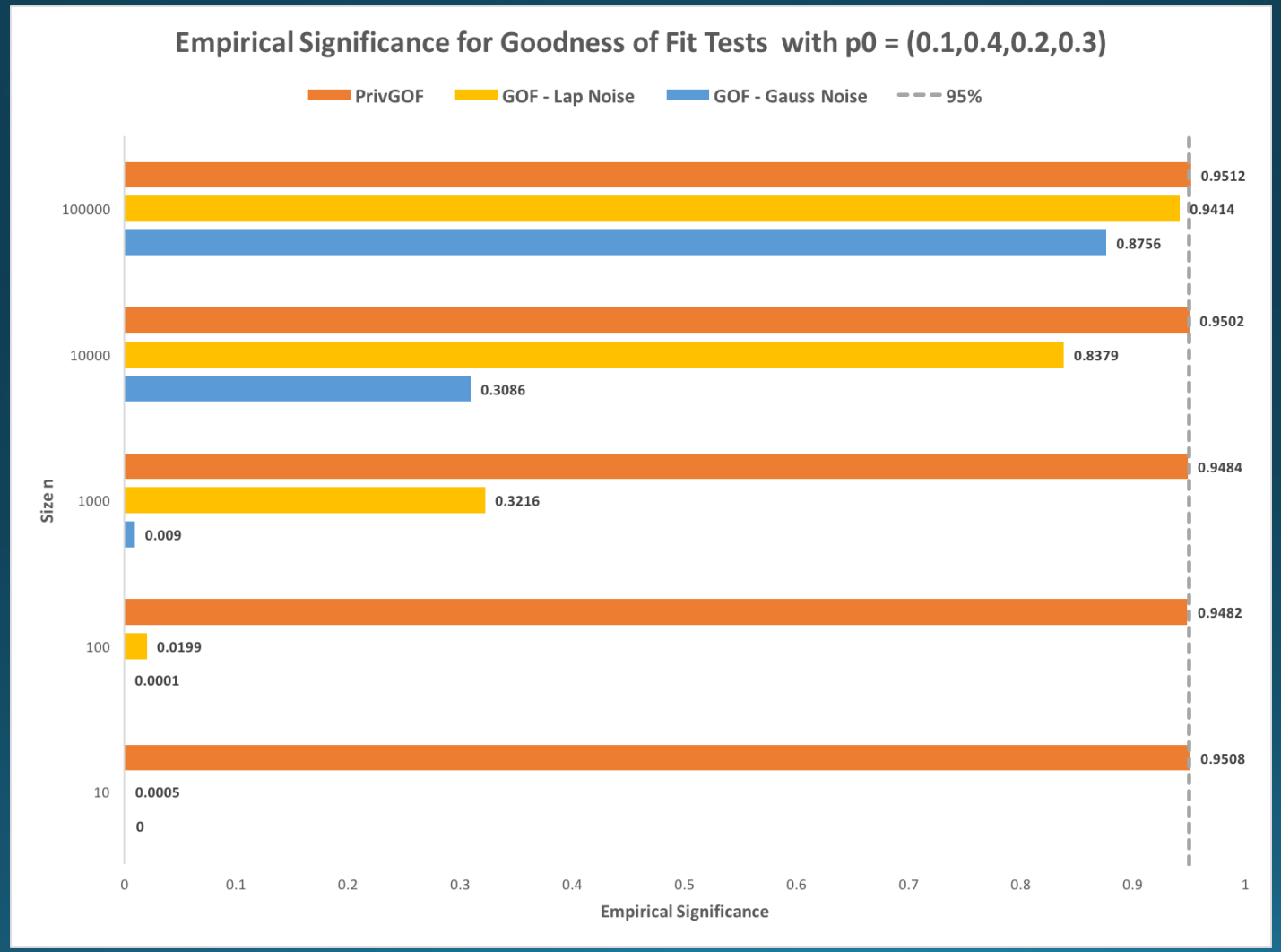
GOF Significance Results

- We fixed the privacy parameters $(\varepsilon, \delta) = (0.1, 10^{-6})$ and $1 - \alpha = 0.95$.
- Sampled 10,000 trials from $H_0 : p = p^0$.
- Counted the proportion of trials that our test did not reject H_0.

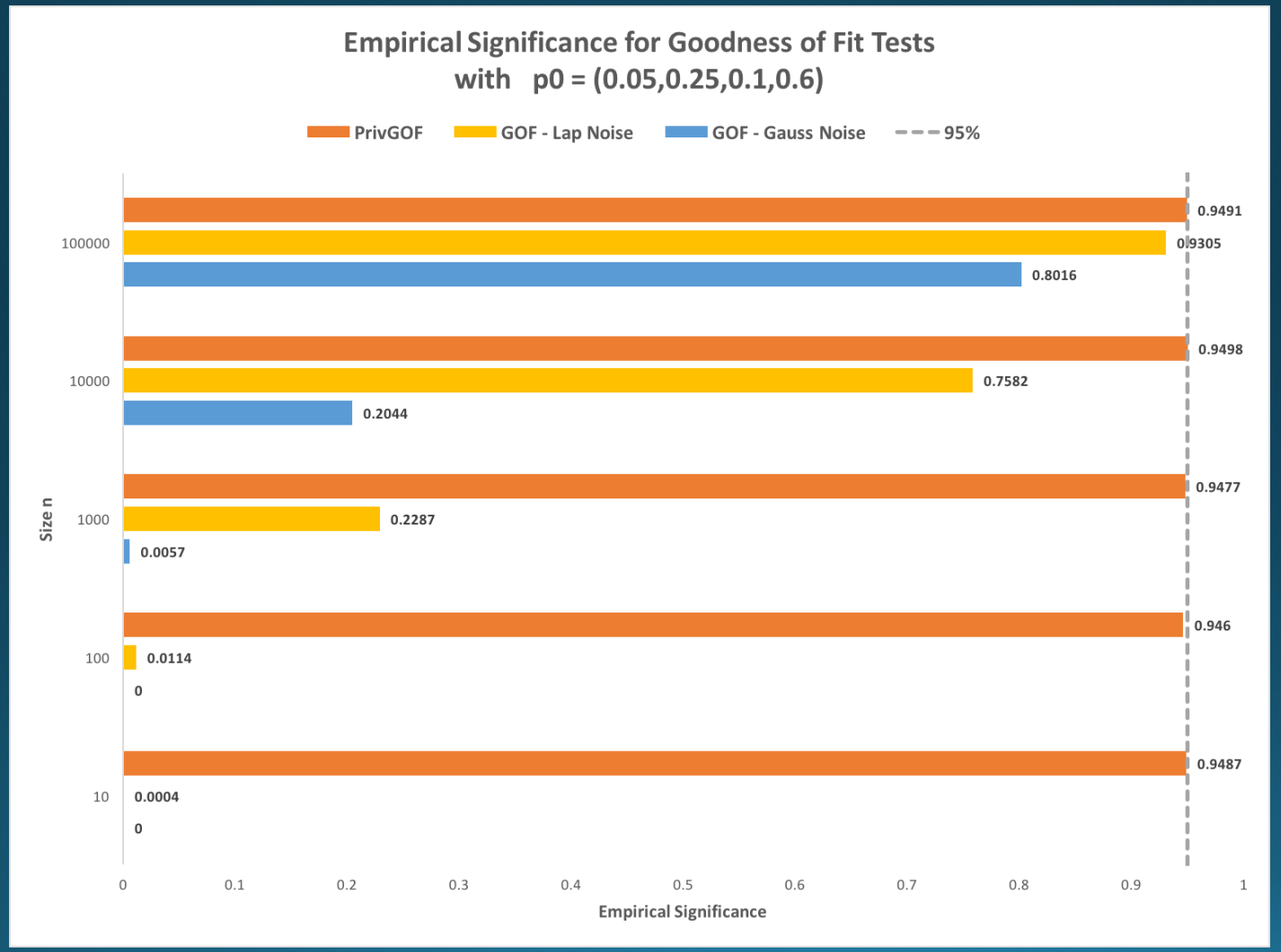
GOF Significance Results



GOF Significance Results



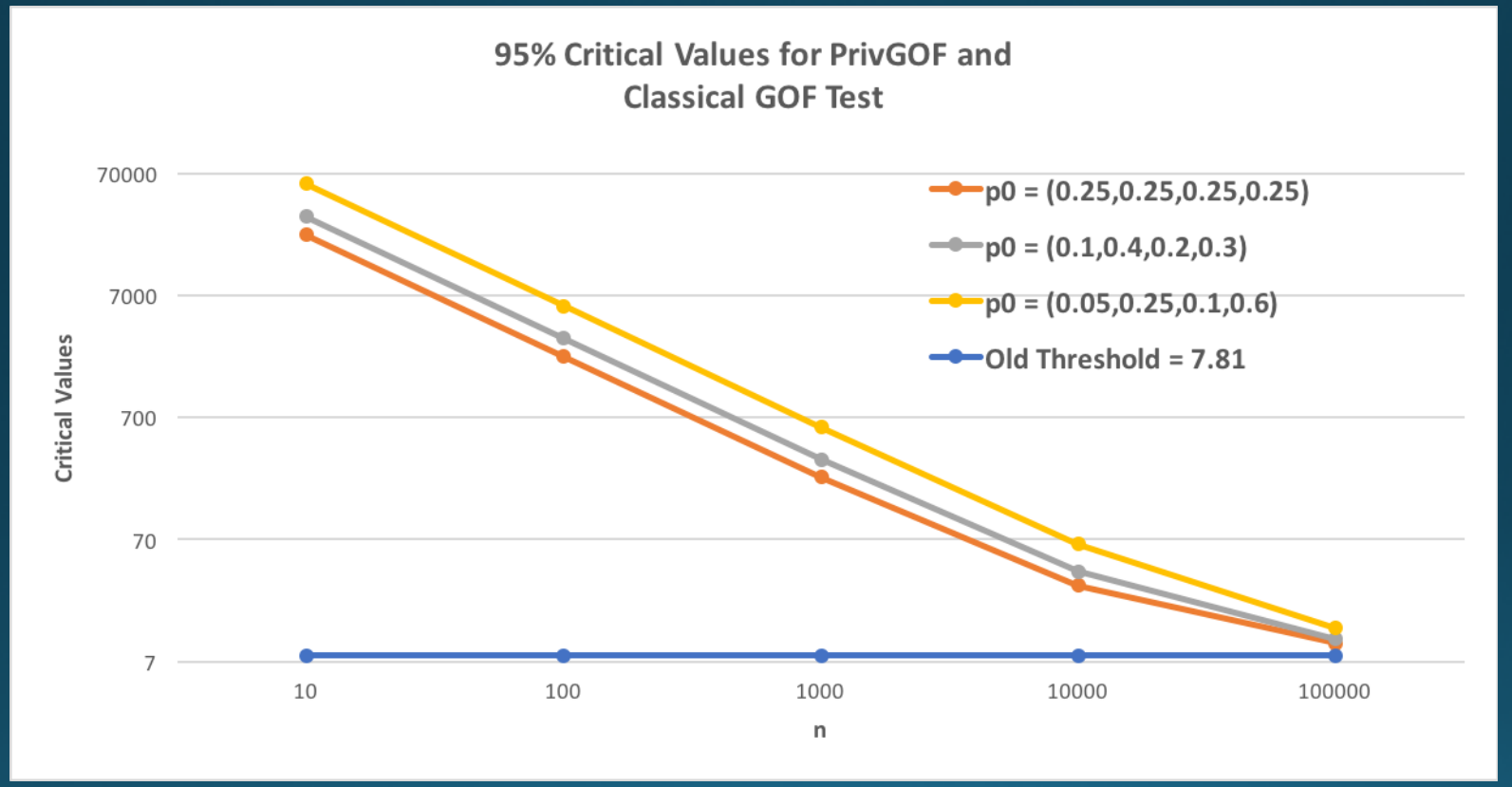
GOF Significance Results



GOF Results with $d = 100$.

$\vec{p}^0$	n	$\chi^2_{d-1,1-\alpha}$	Classical Signif	τ^{α}_{ϵ}	PrivGOF Signif
$\left(\frac{1}{100}, \ldots, \frac{1}{100}\right)$	10,000	123.23	0	7,339	0.9491
$\left(\frac{1}{100}, \ldots, \frac{1}{100}\right)$	100,000	123.23	0	844.7	0.9511
$\left(\frac{1}{100}, \ldots, \frac{1}{100}\right)$	1,000,000	123.23	0.0524	195.3	0.9479

GOF Critical Values - PrivGOF



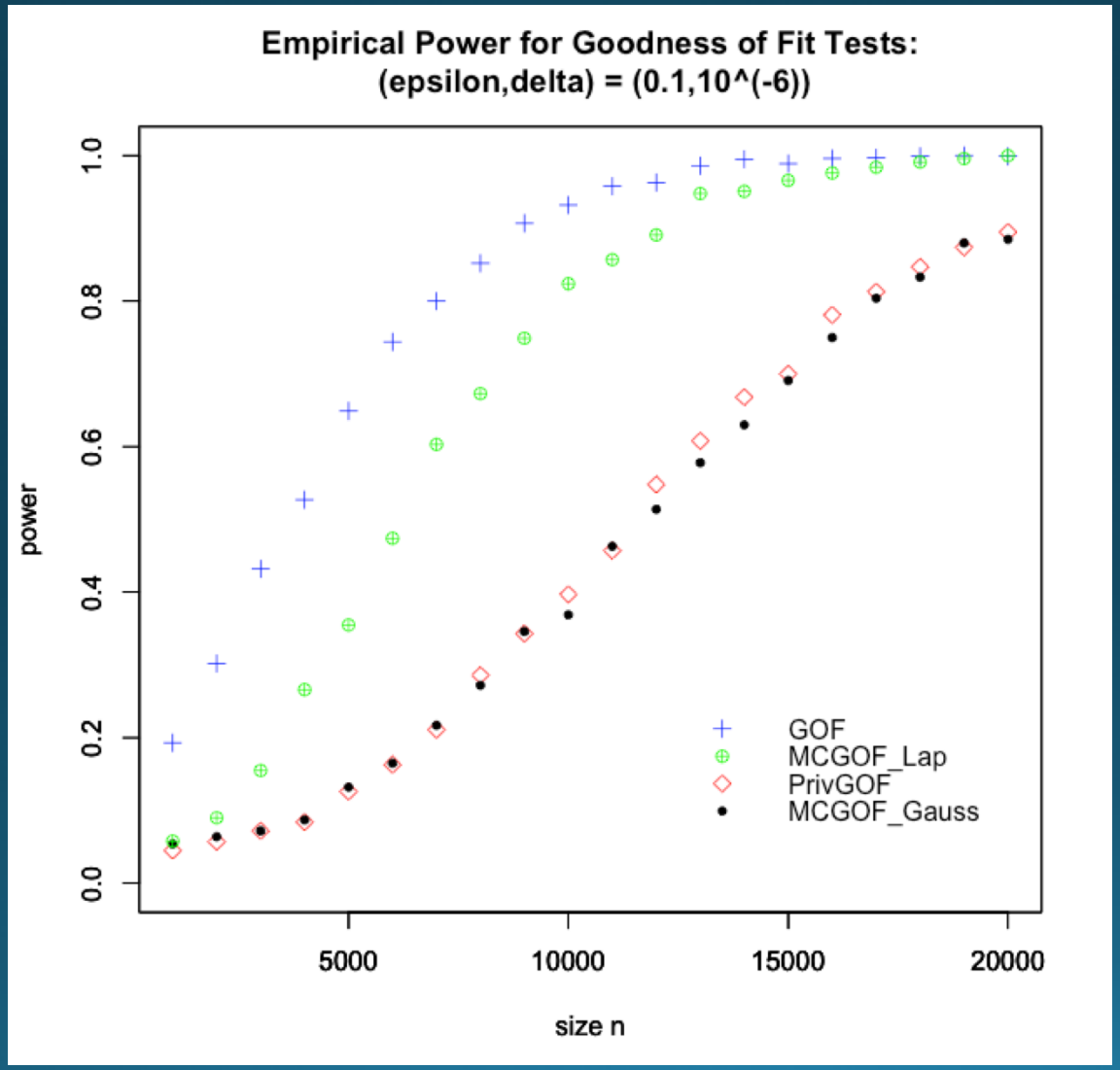
Power of MCGOF and PrivGOF

- Consider the alternate $H_1 : p = p^0 + \Delta \cdot (1, -1, 1, -1, \dots, 1, -1)$.
- Data is actually generated according to H_1 but our test assumes H_0.
- We want our tests to be able to correctly reject H_0 more often as $n \to \infty$.
- We fix the following parameters in our results in 1,000 trials:

$$\alpha = 0.05, (\epsilon, \delta) = (0.1, 10^{-6}), \Delta = 0.01, p^0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

and $k = 50$.

Power Results



Independence Testing: 2 x 2 tables

- Let $Y^{(1)} \sim Bern(\pi^1)$ and $Y^{(2)} \sim Bern(\pi^2)$ and we want to test H_0: $Y^{(1)} \perp Y^{(2)}$.
- Null hypothesis does not completely determine the data generation model - $\pi^1 \; and \; \pi^2$ are unknown.
- Form a contingency table after *n* joint outcomes of $Y^{(1)}$, $Y^{(2)}$

	$Y^{(2)} = 1$	$Y^{(2)} = 0$	**Total**
$Y^{(1)} = 1$	D_{11}	D_{10}	$D_{1\cdot}$
$Y^{(1)} = 0$	D_{01}	D_{00}	$D_{0\cdot}$
Total	$D_{\cdot 1}$	$D_{\cdot 0}$	n

$$D \sim Multinomial(n, p)$$

Under H_0 we have

$$p = \left(\pi^1\pi^2, \pi^1(1-\pi^2), (1-\pi^1)\pi^2, (1-\pi^1)(1-\pi^2)\right)$$

Pearson Chi-Squared Test

- Form the statistic $\hat{Q}^2 = \sum \left[\frac{(D_{ij} - n\widehat{p_{ij}})^2}{n\widehat{p_{ij}}} \right]$
- Using the MLE

$$\widehat{\pi^1} = \frac{D_{1\cdot}}{n}, \widehat{\pi^2} = \frac{D_{\cdot 1}}{n}$$

$$\hat{p} = p\left(\widehat{\pi^1}, \widehat{\pi^2}\right)$$

- Compute df = (rows – 1)(columns – 1)
- If $\hat{Q}^2 > \chi^2_{df,1-\alpha}$, then reject
- Else, fail to reject.

Private MLE

- How do we compute an MLE when we are given private counts?
- Two step procedure to find MLE [inspired by the work of Karwa and Slavkovic '16]
  1. Find most likely true contingency table given the noisy table $D + Z = w$

$$\underset{x}{\operatorname{argmin}} \, ||w - x||$$
$$s.t. \sum_{ij} x_{ij} = n$$
$$x_{ij} \geq 0$$

  2. With this table, compute the MLE for the probability vector as before.

Private MLE

- How do we compute an MLE when we are given private counts?
- Two step procedure to find MLE [inspired by the work of Karwa and Slavkovic '16]
  1. Find most likely true contingency table given the noisy table $D + Z = w$

$$\underset{x}{\operatorname{argmin}} \quad (1-\gamma)||w - x||_1 + \gamma||w - x||_2$$

$$s.t. \sum_{ij} x_{ij} = n$$

$$x_{ij} \geq 0$$

$\gamma = 1$ if Gaussian Noise
$\gamma \ll 1$ if Laplace noise

  2. With this table, compute the MLE for the probability vector as before.

MC Independence Test

- From the noisy contingency table $D + Z$ with Gaussian or Laplace noise, find the private MLE $\tilde{p}$.
- Compute the private chi-squared statistic

$$\tilde{Q}_{DP}^2 = \sum_{ij} \frac{\left(D_{ij} + Z_{ij} - n\tilde{p}_{ij}\right)^2}{n\tilde{p}_{ij}}$$

- With $\tilde{p}$, generate k new contingency tables and add fresh noise to each.
  - From the k new noisy contingency tables, generate k new private chi-squared statistics
  - Set the $\lceil (k+1)(1-\alpha) \rceil -$ ranked value as the critical value τ_ϵ^α.
- If $\tilde{Q}_{DP}^2 > \tau_\epsilon^\alpha$ then reject
- Else, fail to reject.

Asymptotic Approach

- Note that in the Pearson Chi-squared test we have

$\hat{Q}^2 = \hat{U}^T \hat{U}$ where $\hat{U}_{ij} = \frac{D_{ij} - n\hat{p}_{ij}}{\sqrt{n\hat{p}_{ij}}}$ where $\hat{U} \xrightarrow{D} N(0, \Sigma_{ind})$

- Now Σ_{ind} depends on the unknown probabilities
  - (not the same as Σ from before).
  - $\Sigma_{ind} = I - \sqrt{p}\sqrt{p}^T - \Gamma(\Gamma^T\Gamma)^{-1}\Gamma^T$

Asymptotic Approach

- We will follow the same procedure as in the GOF testing, except we will use the private MLE $\tilde{p}$ whenever we would have used the actual probability vector.

$$\tilde{W} = (\tilde{U}, V), \qquad \tilde{Q}_{DP}^2 = \tilde{W}^T \tilde{A} \tilde{W}$$

- Where with Guassian noise we have,

$$\tilde{W} \approx N(0, \Sigma'_{ind}), \qquad \Sigma'_{ind} = \begin{bmatrix} \Sigma_{ind} & 0 \\ 0 & I \end{bmatrix}$$

- We will use $\widetilde{\Sigma'}_{ind}$ which just replaces the unknown probability vector in Σ'_{ind} with our private MLE $\tilde{p}$.

PrivInd

- We approximate the distribution of $\tilde{Q}_{DP}^2$ with

$$\sum_i \tilde{\lambda}_i \chi_1^2,$$

where $\{\tilde{\lambda}_i\}$ are the eigenvalues of $\tilde{B}^T \tilde{A} \tilde{B}$ and $\tilde{B}\tilde{B}^T = \tilde{\Sigma}'_{ind}$.

- New Test - PrivInd:
  - Compute the private MLE $\tilde{p}$ based on noisy counts
  - Compute the statistic $\tilde{Q}_{DP}^2$ and critical value τ_ϵ^α where,

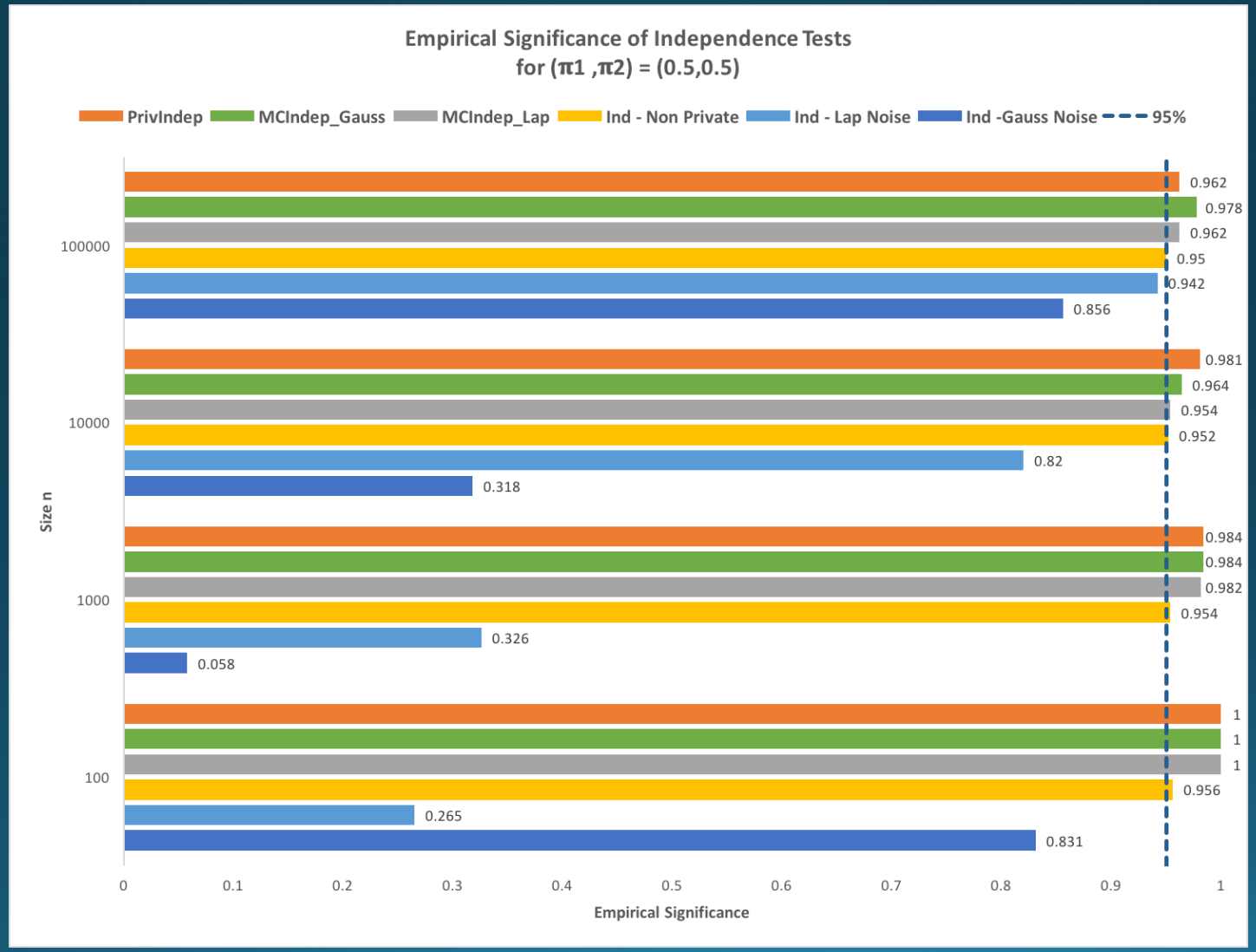
$$P\left[\sum_i \tilde{\lambda}_i \chi_1^2 > \tau_\epsilon^\alpha\right] = \alpha$$

  - If $\tilde{Q}_{DP}^2 > \tau_\epsilon^\alpha$, reject.
  - Else, fail to reject.

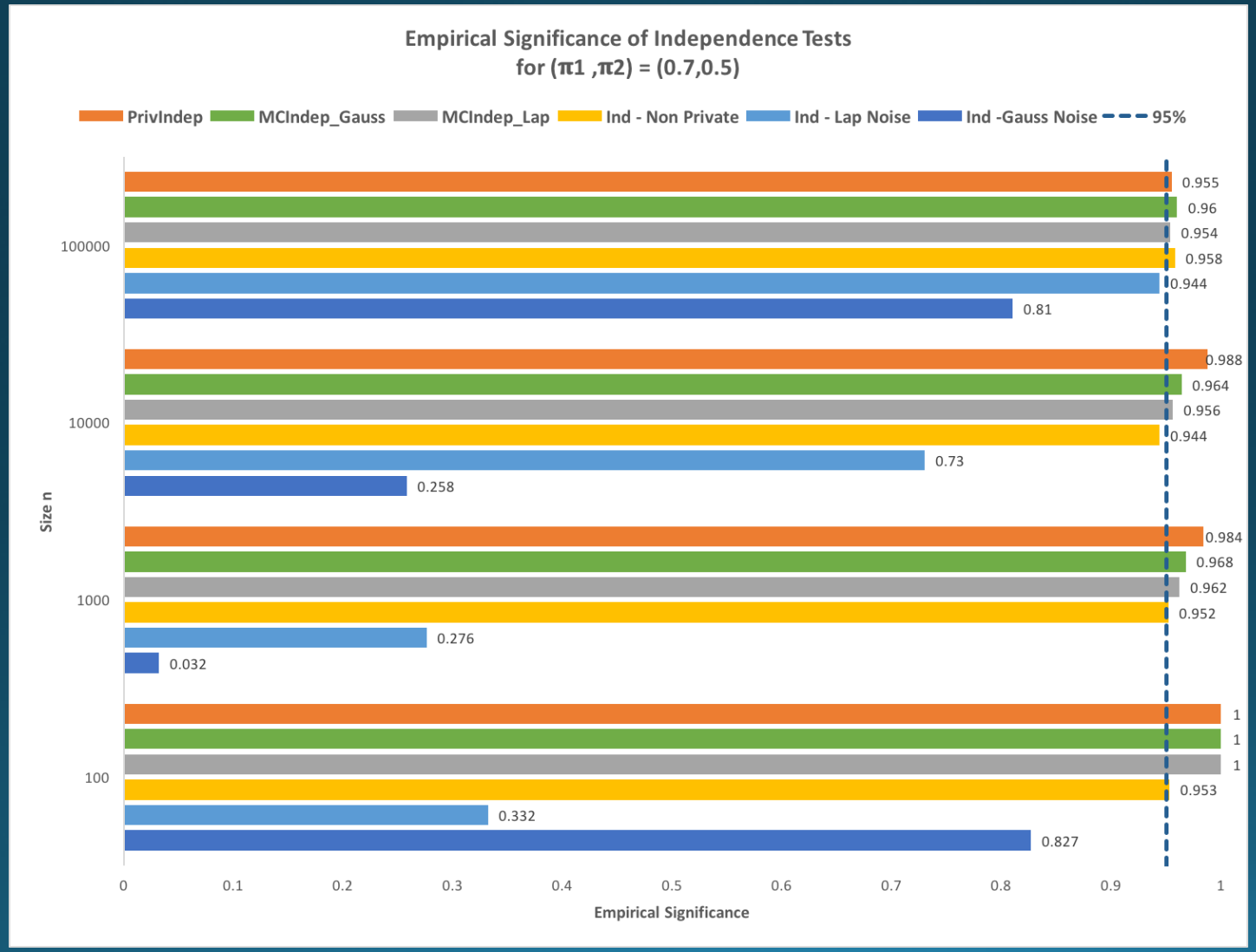
Independence Significance Results

- We fixed the privacy parameters $(\varepsilon, \delta) = (0.1, 10^{-6})$ and $1 - \alpha = 0.95$.
- Sampled 1,000 trials of independent data $Y^{(1)} \sim Bern(\pi^1)$ and $Y^{(2)} \sim Bern(\pi^2)$ for various values of π^1, π^2.
- Counted the proportion of trials that our test did not reject H_0.

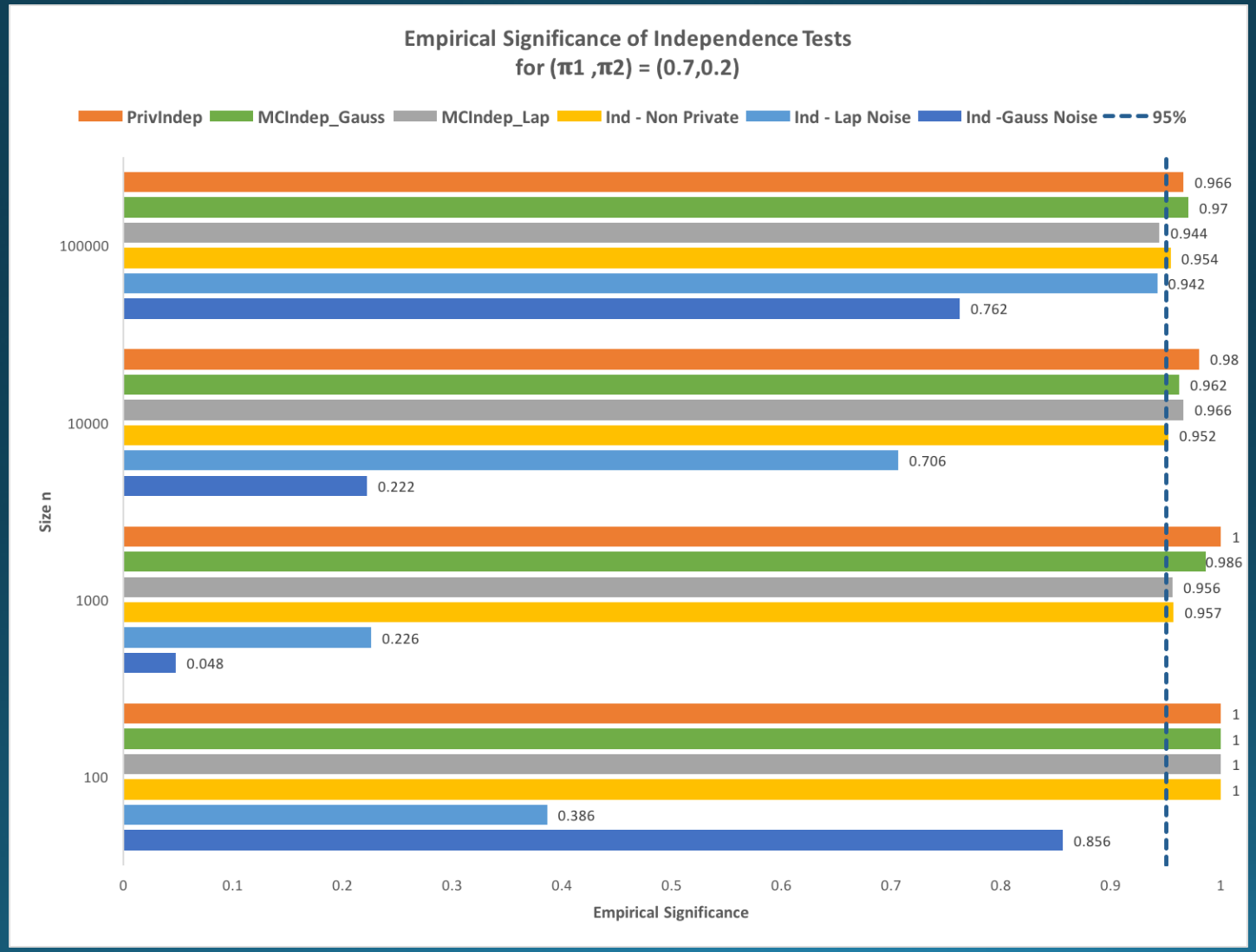
Independence Significance Results



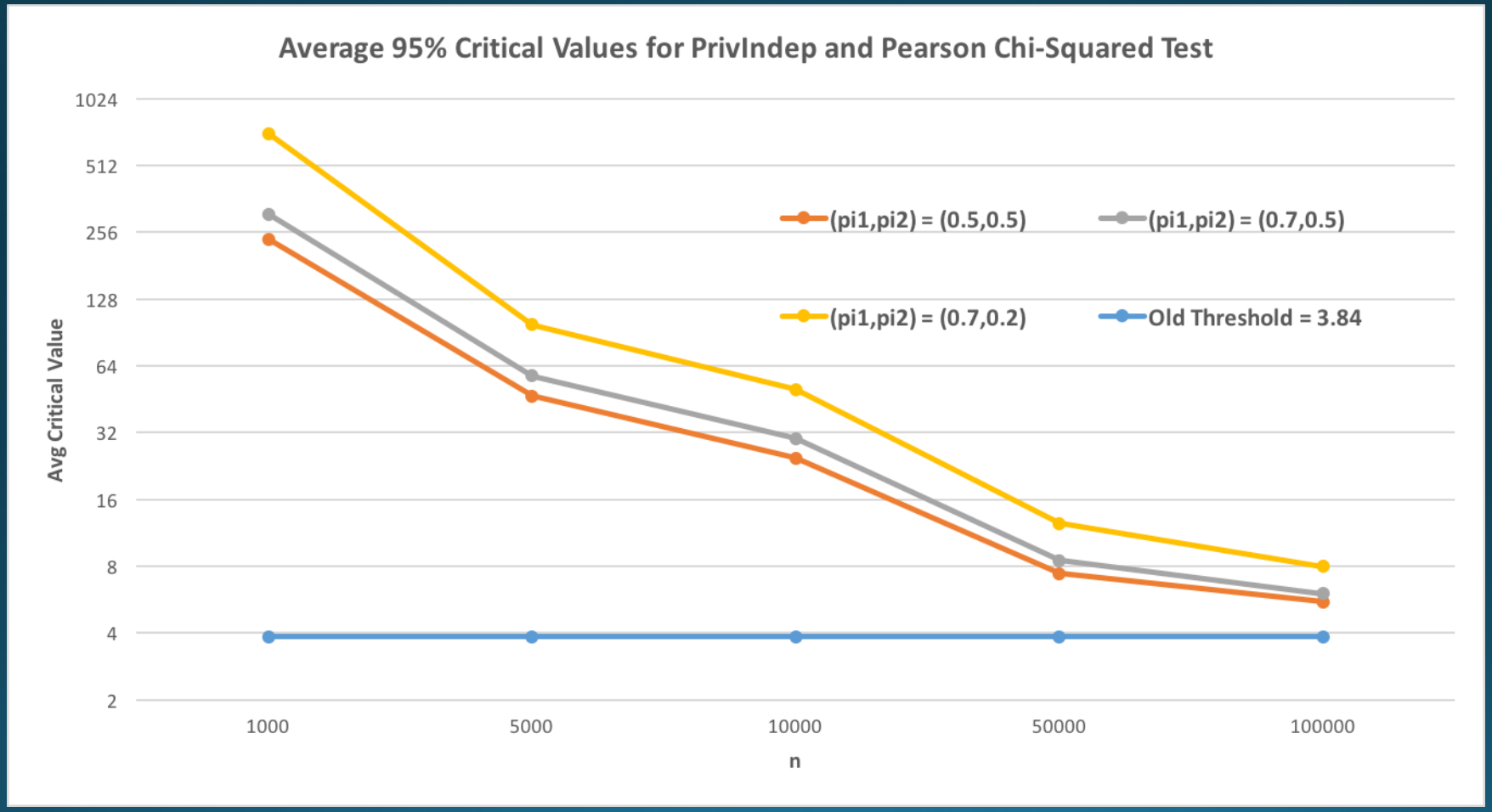
Independence Significance Results



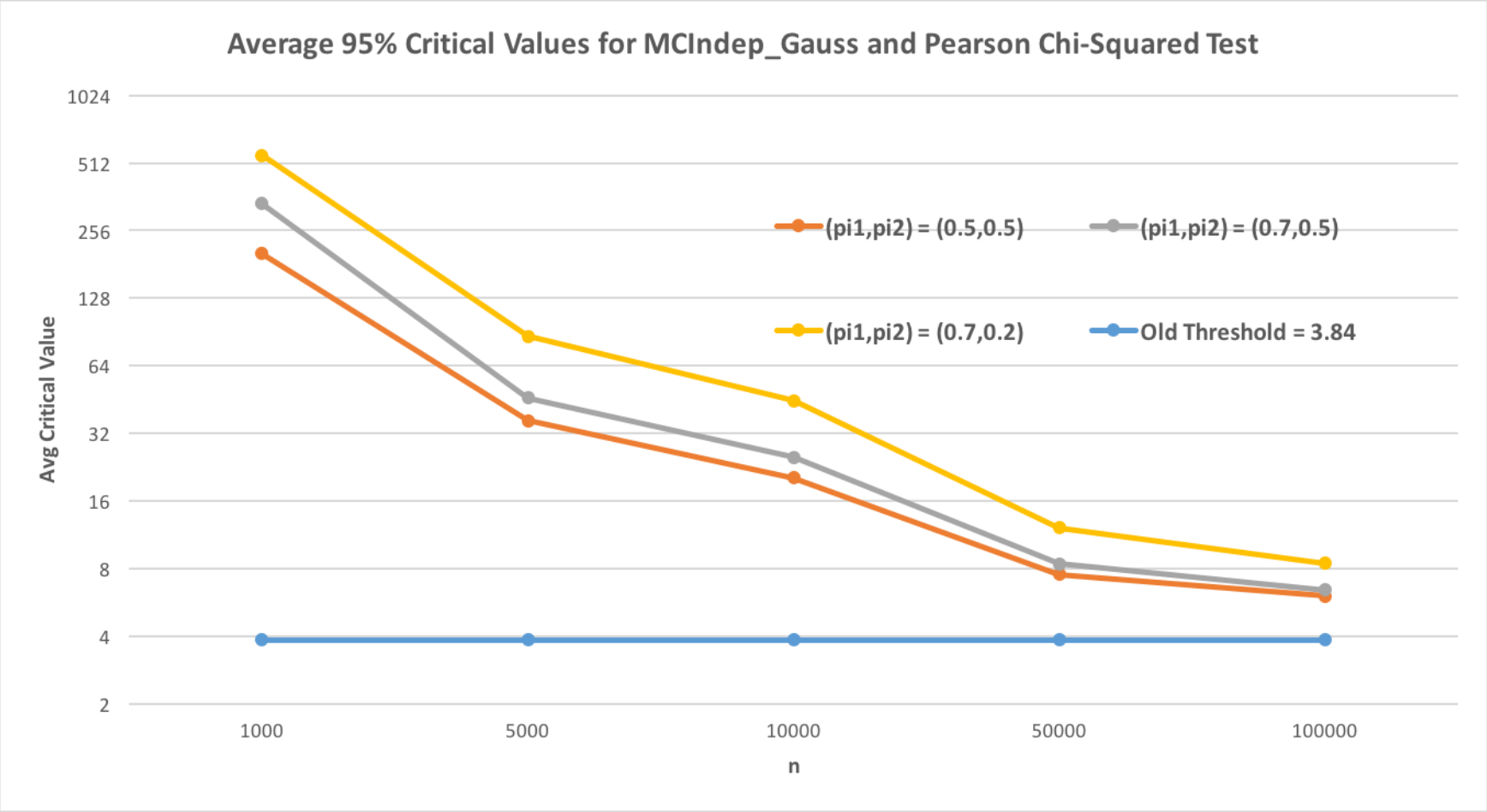
Independence Significance Results



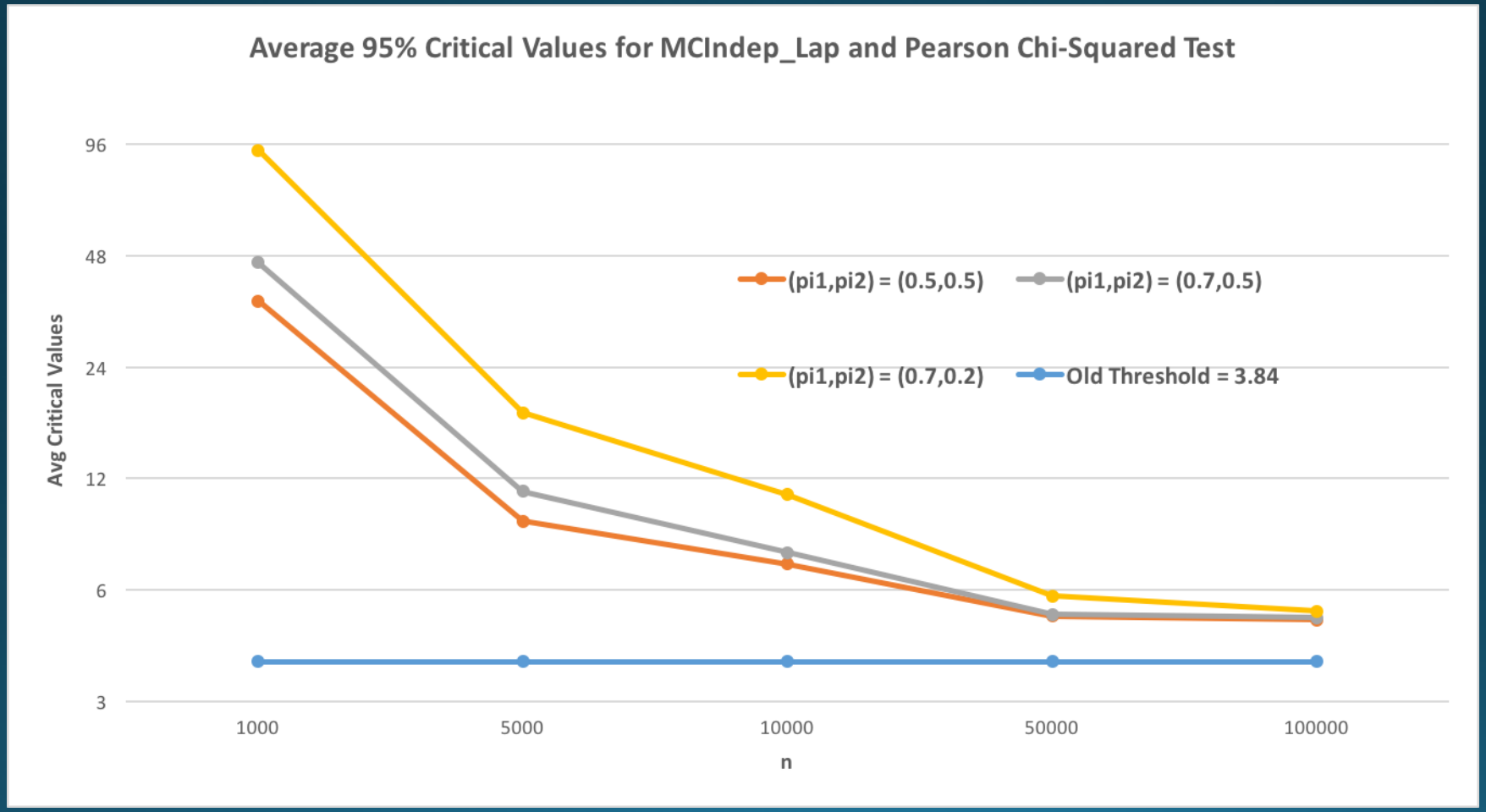
Critical Values of PrivInd



Critical Values of MCIndep_Gauss



Critical Values of MCIndep_Lap



Testing Power

- Consider the alternate hypothesis $H_1 : Cov(Y^{(1)}, Y^{(2)}) = \Delta$.
- The table of counts then come from the following distribution:

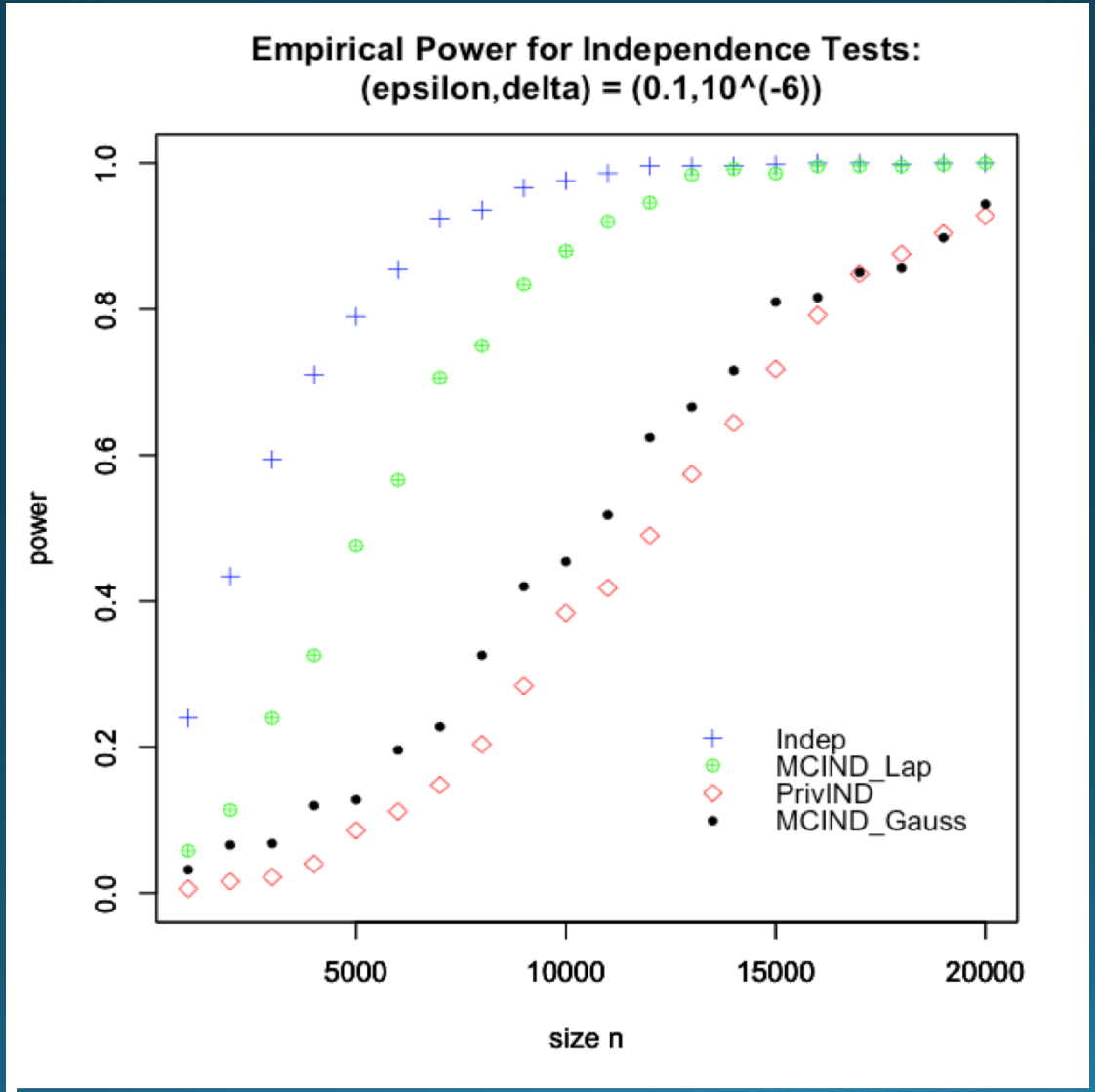
$$D \sim Multinomial(n, p + \Delta(1, -1, 1, -1)$$

Where $p = \left(\pi^1\pi^2, \pi^1(1-\pi^2), (1-\pi^1)\pi^2, (1-\pi^1)(1-\pi^2)\right)$

- We then use the parameters:

$\alpha = 0.05,$ $(\epsilon, \delta) = (0.1, 10^{-6}),$ $\Delta = 0.01,$ $(\pi^1, \pi^2) = \left(\frac{1}{2}, \frac{1}{2}\right)$

Power Results



Conclusion

- Developed four DP tests with at least $1 - \alpha$ empirical significance:
  - GOF Testing
    1. MCGOF – Gaussian or Laplace noise with guaranteed significance at least $1 - \alpha$
    2. PrivGOF – Only works with Gaussian noise, based on asymptotic approach
  - Independence Testing
    1. MCInd – Gaussian or Laplace Noise
    2. PrivInd – Only works with Gaussian noise, based on asymptotic approach
- Tests based on Laplace noise have better power, due to the smaller variance of the noise being added to the counts.
- Laplace noise tests rely on MC methods.
- PrivGOF and PrivInd resemble the classical tests.

Thanks