A New Class of Private Chi-Square Tests

Daniel Kifer, Ryan Rogers

PENNSTATE 1855

Penn UNIVERSITY of PENNSYLVANIA

Hypothesis Testing

	H_0 True	H_0 False
Reject H_0	False Discovery	Power
Not	Significance	Type II Error

- Given dataset D and proposed model of H_0, should H_0 should be rejected or not based on data.
- **Goal**: Bound $\mathbb{P}$[False Discovery] $\leq \alpha$, while obtaining good power.

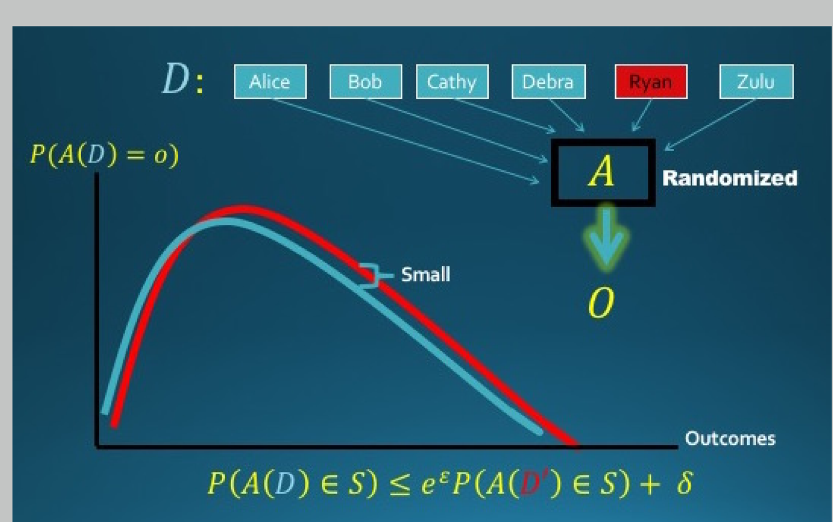
The Need for Privacy

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

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

(Concentrated) Differential Privacy [DMNS], [BS]

- Outcome of test $A : \mathcal{D} \to \mathcal{O}$ should *roughly* stay the same if one person changes his data.
- DP [DMNS]: For any neighboring D, D' and outcome set $S \subseteq \mathcal{O}$:
$$\mathbb{P}[A(D) \in S] \leq e^{\epsilon}\mathbb{P}[A(D') \in S] + \delta.$$
- zCDP [BS]: Another measure of privacy "between" $(\epsilon, 0)$-DP and $(\epsilon, \delta > 0)$-DP.



Focus of this work: Chi-Square Tests

- Categorical data histogram: $X \sim$ Multinomial$(n, \mathbf{p} = (p_1, \cdots, p_d))$
- General class of tests use 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**: $H_0 : Y^1 \sim$ Multinomial$(1, \pi^1)$ and $Y^2 \sim$ 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

- [USF13, YFSU14] Add noise to statistic to preserve privacy $\to$ leads to unbounded noise in worst case.
- [JS] Add noise to histogram, use classical test $\to$ leads to $\mathbb{P}$[False Discovery] $> \alpha$ for small datasets.
- [GLRV, WLK15] Add noise to histogram, use classical statistic but modify distribution to take into account the noise.
- This Work: Add noise to histogram, modify statistic to account for the noise so that it is a chi-square random variable as in the classical tests.

Preliminaries

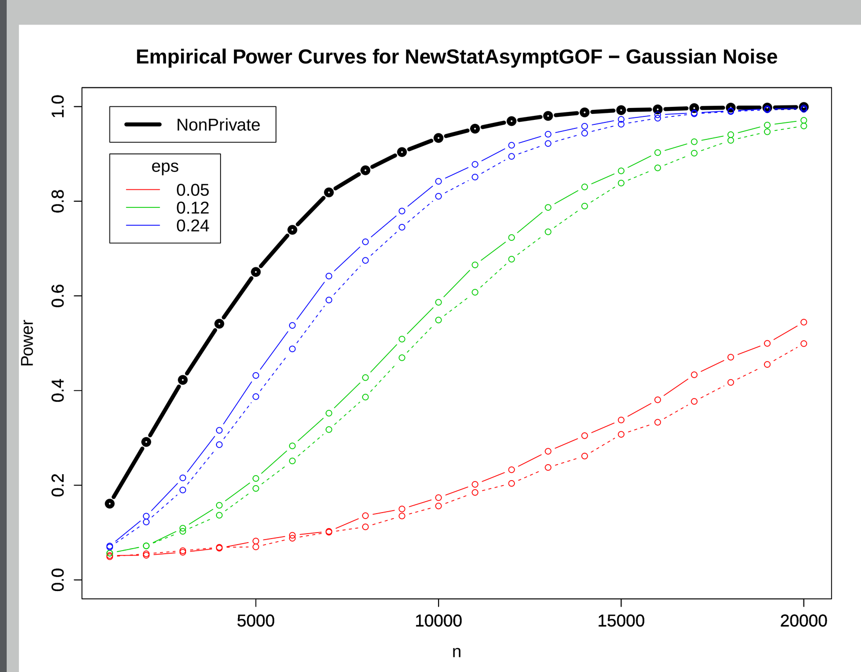
- Add $N(0, \sigma^2)$ noise to each cell count of histogram, get noisy version $\tilde{X}$
- Write covariance matrix for multinomial with added noise
$$\Sigma_{DP} = \mathtt{Diag}(\mathbf{p}^0) - \mathbf{p}^0\left(\mathbf{p}^0\right)^\intercal + \sigma^2 I_d$$

Two New Test Statistics for DP Hypothesis Tests

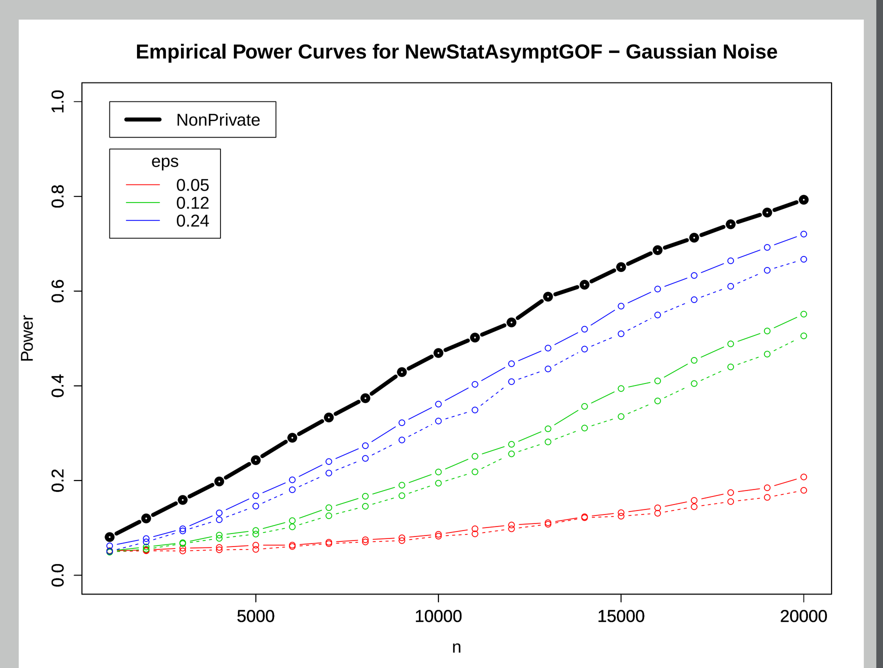
- The *unprojected* statistic:
$$Q^2_{DP} = n\left(\tilde{X}/n - \mathbf{p}^0\right)^\intercal \Sigma^{-1}_{DP}\left(\tilde{X}/n - \mathbf{p}^0\right)$$
- **Theorem:** Under the null hypothesis and $\sigma^2/n \to$ constant > 0,
$$Q^2_{DP} \xrightarrow{D} \chi^2_d.$$
- The *projected* statistic with projection $\Pi = I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\intercal$:
$$\mathcal{Q}^2_{DP} = n\left(\tilde{X}/n - \mathbf{p}^0\right)^\intercal \Pi\Sigma^{-1}_{DP}\Pi\left(\tilde{X}/n - \mathbf{p}^0\right)$$
- **Theorem:** Under the null hypothesis and $\sigma^2 = O(n)$,
$$\mathcal{Q}^2_{DP} \xrightarrow{D} \chi^2_{d-1}.$$
- Difference of two statistics is a scaled *independent* chi-square
$$Q^2_{DP} - \mathcal{Q}^2_{DP} = d\,\sigma^2\,\chi^2_1$$
- Statistics can be extended to more general chi-square tests and we can use other types of noise distributions via an MC approach.

Power Results – also works with Laplace noise

- Test is designed to achieve $\mathbb{P}$[False Discovery] close to α, as in classical test.
- Experimentally check the power of our test in 10,000 trials, fixing $\delta = 10^{-6}$ (projected stat is solid and unprojected is dashed)

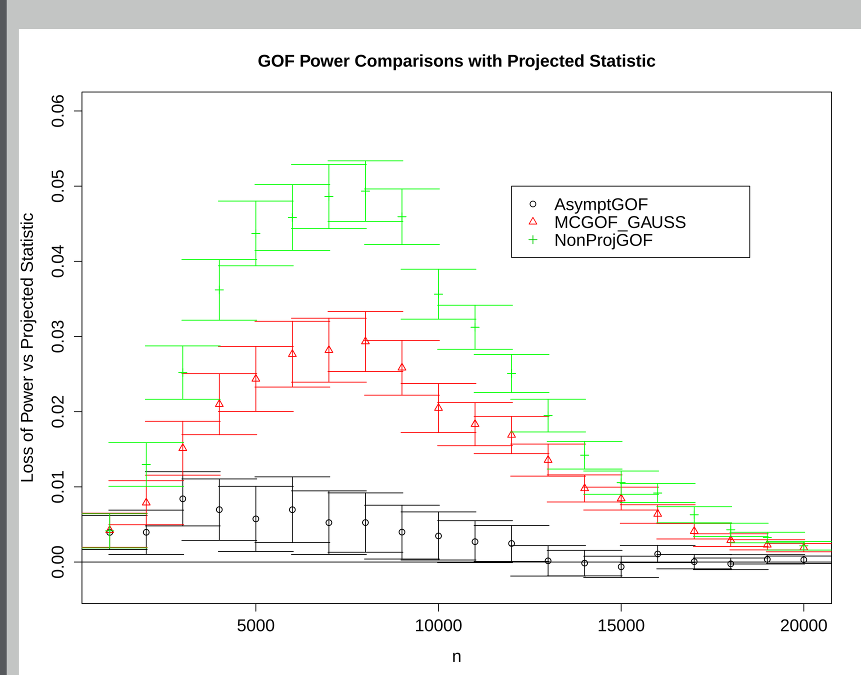


$H_0 : \mathbf{p}^0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})^\intercal$,
$H_1 : \mathbf{p}^0 + 0.01 \cdot (1, -1, -1, 1)^\intercal$

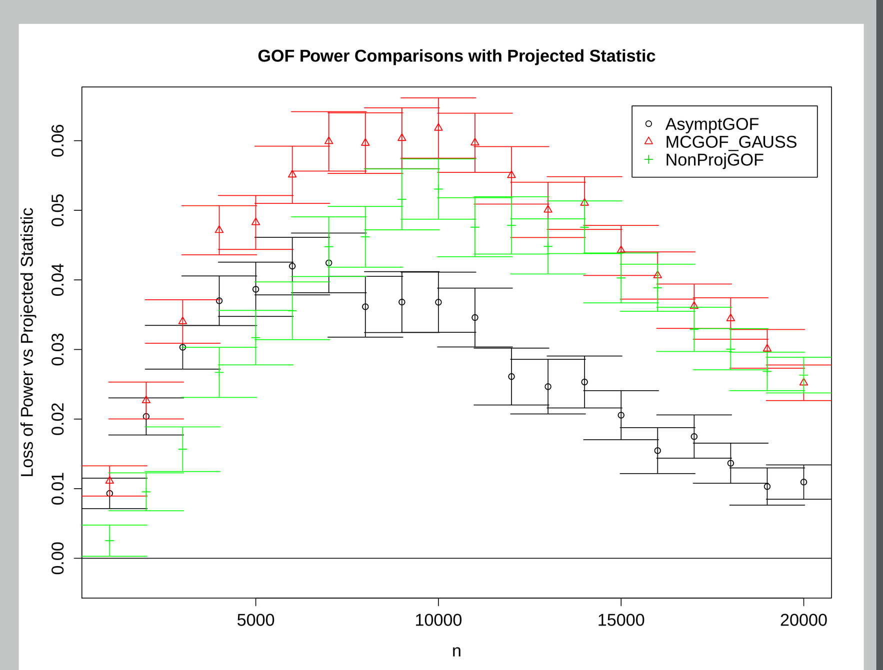


$H_0 : \mathbf{p}^0 = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})^\intercal$,
$H_1 : \mathbf{p}^0 + 0.01 \cdot (1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})^\intercal$.

Power Comparison with [GLRV]: $\alpha = 0.05$, $(\epsilon, \delta) = (0.24, 10^{-6})$



$H_0 : \mathbf{p}^0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})^\intercal$,
$H_1 : \mathbf{p}^0 + 0.01 \cdot (1, -1, -1, 1)^\intercal$



$H_0 : \mathbf{p}^0 = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})^\intercal$,
$H_1 : \mathbf{p}^0 + 0.01 \cdot (1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})^\intercal$.

References

[BS] Bun and Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC'16,Part I*.

[DMNS] Dwork, McSherry, Nissim, and Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*.

[GLRV] Gaboardi, Lim, Rogers, and Vadhan. Differentially private chi-squared hypothesis testing. In *ICML'16*.

[JS] Johnson and Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *KDD'13*.

[USF13] Uhler, Slavkovic, and Fienberg. Privacy-preserving data sharing for gwas. *J. of Privacy and Confidentiality*, 5(1), 2013.

[WLK15] Wang, Lee, and Kifer. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015.

[YFSU14] Yu, Fienberg, Slavković, and Uhler. Scalable privacy-preserving data sharing methodology for gwas. *J. of Biomed Informatics*, 50, 2014.