# Local Private Hypothesis Tests: Chi-Square Tests

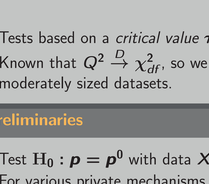
Marco Gaboardi, Ryan Rogers

## Hypothesis Testing

| | $\mathbf{H_0}$ True | $\mathbf{H_0}$ False |
|---|---|---|
| Reject $\mathbf{H_0}$ | False Discovery | Power |
| Not | Significance | Type II Error |

- Given dataset and proposed model of $\mathbf{H_0}$, should it 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.

## Local Differential Privacy [DMNS], [RSL$^+$]

- Central Model: Data is submitted in the clear to a trusted curator and the output of a statistic on the data is privatized.
- Local Model: No trusted curator - data is privatized and then collected.
- An algorithm $M : \mathcal{X} \to \mathcal{O}$ is $\epsilon$-differentially private if for all inputs, $x$, $x'$ and outcome sets $S \subseteq \mathcal{O}$:

$$\mathbb{P}[M(x) \in S] \leq e^{\epsilon}\mathbb{P}[M(x') \in S].$$

- Local model of differential privacy is used in practice.

## Focus of this work: Chi-Square Tests

- Categorical data entries histogram: $X_i \sim \text{Multinomial}(1, \mathbf{p} = (p_1, \cdots, p_d))$
- General class of tests use the chi-square statistic based on histogram $H = \sum_i X_i$:

$$Q^2 = \sum_j \frac{(\text{Observed}[j] - \text{Expected}[j])^2}{\text{Expected}[j]}.$$

- **Goodness of Fit**: $\mathbf{H_0 : p = p^0}$.
- **Independence Testing**: $\mathbf{H_0} : Y^{(1)} \sim \text{Multinomial}(1, \pi^{(1)})$ 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* $\tau$, so that if $Q^2 > \tau$ then reject $\mathbf{H_0}$.
- 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 $\alpha$. Works well even for moderately sized datasets.

## Preliminaries

- Test $\mathbf{H_0 : p = p^0}$ with data $X_1, \cdots X_n \sim \text{Multinomial}(\mathbf{1}, \boldsymbol{p})$
- For various private mechanisms $M$, let $H = \sum_{i=1}^n M(X_i)$
- We use the technique from [KR] to reduce the number of degrees of freedom for the asymptotic distribution of the test statistic.
- Define the projection $\mathbf{\Pi} = \left(I_d - \frac{1}{d} \cdot \mathbf{1}\mathbf{1}^{\intercal}\right)$ and covariance matrix $\mathbf{\Sigma^0} = \text{Diag}(\boldsymbol{p^0}) - \boldsymbol{p^0}\left(\boldsymbol{p^0}\right)^{\intercal}$.

## Prior Work for DP Hypothesis Tests — All in Central Model

- [USF, YFSU] Add noise to statistic to preserve privacy $\rightarrow$ leads to unbounded noise in worst case.
- [JS] Add noise to histogram, use classical test $\rightarrow$ leads to $\mathbb{P}$ [False Discovery] $> \alpha$ for small datasets.
- [GLRV, WLK] Add noise to histogram, use classical statistic but modify distribution to take into account the noise.
- [KR]: 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.

## Local Private Test Statistics

**Test statistics for various private mechanisms**

- `LocalNoiseGOF`: Add noise to data $M(X_i) = X_i + Z_i$ , s.t. $\mathbb{V}[Z_i] = \sigma^2$

$$\mathrm{T}_{\text{Noise}} = n\left(\frac{H}{n} - \boldsymbol{p^0}\right)^{\intercal} \mathbf{\Pi}\left(\mathbf{\Sigma^0} + \sigma^2 \cdot I_d\right)^{-1} \mathbf{\Pi}\left(\frac{H}{n} - \boldsymbol{p^0}\right)$$

- `LocalGenRRGOF`: Use generalized randomized response, i.e. $M(X_i) = X_i$ with probability $\frac{e^{\epsilon}}{e^{\epsilon}+d-1}$, otherwise $M(X_i) \neq X_i$. Set $\tilde{\boldsymbol{p}}^{\mathbf{0}} = \left(\frac{e^{\epsilon}-1}{e^{\epsilon}+d-1}\right) \cdot \boldsymbol{p^0} + \frac{1}{e^{\epsilon}+d-1}$

$$\mathrm{T}_{\text{RR}} = n\left(\frac{H}{n} - \tilde{\boldsymbol{p}}^{\mathbf{0}}\right)^{\intercal} \text{Diag}\left(\tilde{\boldsymbol{p}}^{\mathbf{0}}\right)^{-1}\left(\frac{H}{n} - \tilde{\boldsymbol{p}}^{\mathbf{0}}\right)$$

- `LocalBitFlipGOF`: Using a mechanism from [BS] we use $M(X_i) = (M_1(X_i[1]), \cdots, M_d(X_i[d]))$ where $M_i(b) = b$ w.p $\frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$ and $M(b) = 1 - b$ otherwise. Set $\kappa = \frac{e^{\epsilon/2}-1}{e^{\epsilon/2}+1}$ and $\check{\boldsymbol{p}}^{\mathbf{0}} = \kappa \cdot \boldsymbol{p^0} + \frac{1}{e^{\epsilon/2}+1}$

$$\mathrm{T}_{\text{BF}} = \frac{n}{\kappa^2}\left(\frac{H}{n} - \check{\boldsymbol{p}}^{\mathbf{0}}\right)^{\intercal} \mathbf{\Pi}\left(\mathbf{\Sigma^0} + \frac{e^{\epsilon/2}}{\left(e^{\epsilon/2}-1\right)^2} \cdot I_d\right)^{-1}\left(\frac{H}{n} - \check{\boldsymbol{p}}^{\mathbf{0}}\right)$$

**All statistics converge in distribution to $\chi^2_{d-1}$ under $\mathbf{H_0}$.**

## Noncentral Parameters Comparison

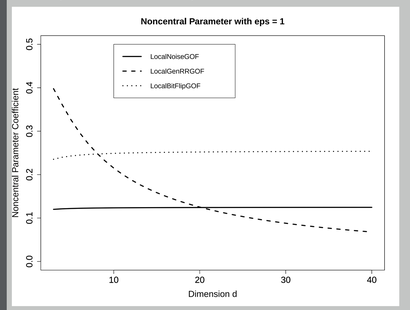
- Each test is designed to achieve $\mathbb{P}$[False Discovery] at most $\alpha$ asymptotically, as in classical test.
- Test: $\mathbf{H_0 : p^0} = (1/d, \cdots, 1/d)$. $\mathbf{H_1 : p^1} = \mathbf{p^0} + \frac{1}{\sqrt{n}}\mathbf{\Delta}$.
- Under $\mathbf{H_1}$, we have

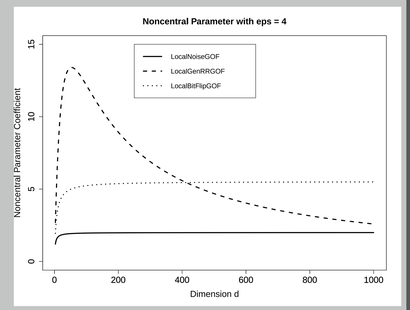
$$\mathrm{T}_{\text{Noise}} \xrightarrow{D} \chi^2_{d-1}\left(\left(\frac{d}{1+d\sigma^2}\right) \cdot ||\mathbf{\Delta}||_2^2\right), \qquad \sigma \approx 1/\epsilon$$

$$\mathrm{T}_{\text{RR}} \xrightarrow{D} \chi^2_{d-1}\left(d \cdot \left(\frac{e^{\epsilon}-1}{e^{\epsilon}+d-1}\right)^2 \cdot ||\mathbf{\Delta}||_2^2\right)$$

$$\mathrm{T}_{\text{BF}} \xrightarrow{D} \chi^2_{d-1}\left(d \cdot \left(\frac{(e^{\epsilon/2}-1)^2}{d \cdot e^{\epsilon/2} + \left(e^{\epsilon/2}-1\right)^2}\right) \cdot ||\mathbf{\Delta}||_2^2\right)$$

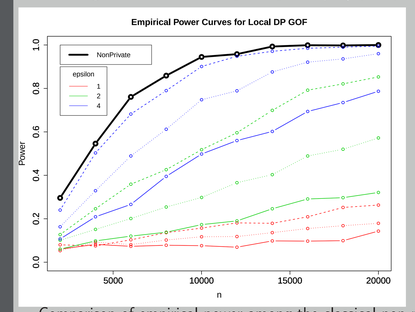
- We plot the noncentral parameter for the various test statistics,

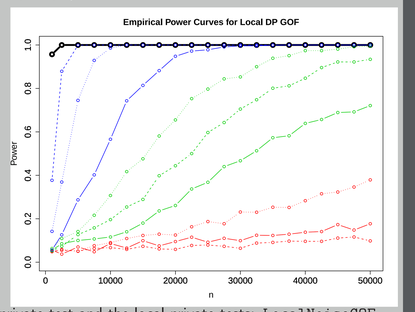



## Power Comparison for Goodness of Fit Testing

- Test: $\mathbf{H_0 : p^0} = (1/d, \cdots, 1/d)$.
- Data is generated from $\mathbf{H_1 : p^1} = \mathbf{p^0} + \eta \cdot (1, -1, \cdots, -1, 1)$.

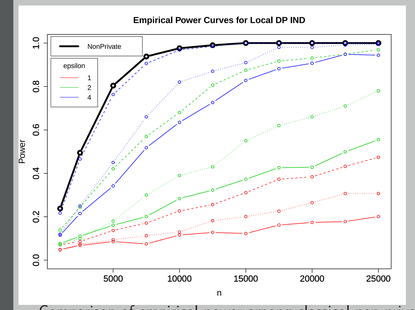


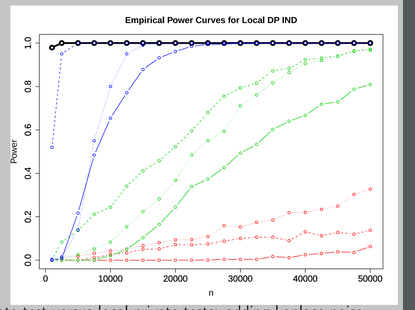




Comparison of empirical power among the classical non-private test and the local private tests: `LocalNoiseGOF` with Laplace noise (solid line), `LocalGenRRGOF` (dashed line), and `LocalBitFlipGOF` (dotted line); in the left plot $d = 4$ and $\eta = 0.01$, in the right plot $d = 40$ and $\eta = 0.005$.

## Power Comparison for Independence Testing

- Test $\mathbf{H_0} : Y^{(1)} \perp Y^{(2)}$ with contingency table data $\{X_{j,\ell} : j \in [r], \ell \in [c]\}$.
- Data generated with $\boldsymbol{\pi^{(1)}}\left(\boldsymbol{\pi^{(2)}}\right)^{\intercal} + \eta \cdot (1, -1, \cdots, -1, 1)^{\intercal}(1, -1, \cdots, -1, 1)$ with symmetric $\boldsymbol{\pi^{(i)}}$.




Comparison of empirical power among classical non-private test versus local private tests: adding Laplace noise (solid line), `LocalGenRRIND` (dashed line), and `LocalBitFlipIND` (dotted line) where the left plot $(r, c) = (2, 2)$ and $\eta = 0.01$, the right plot $(r, c) = (10, 4)$ and $\eta = 0.005$

## References


[BS] Bassily and Smith. Local, private, efficient protocols for succinct histograms. In *STOC'15*.

[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*.

[KR] Kifer and Rogers. A New Class of Private Chi-Square Hypothesis Tests. In *AISTATS'17*.

[RSL$^+$] Raskhodnikova, Smith, H. K. Lee, Nissim, and Kasiviswanathan. What can we learn privately? *IEEE'08*.

[USF] Uhler, Slavkovic, and Fienberg. Privacy-preserving data sharing for gwas. *J. of Privacy and Confidentiality'13*.

[WLK] Wang, J. Lee, and Kifer. Differentially private hypothesis testing, revisited. *arXiv:1511.03376, '15*.

[YFSU] Yu, Fienberg, Slavković, and Uhler. Scalable privacy-preserving data sharing methodology for gwas. *J. of Biomed Informatics'14*, 50.