Differentially Private Estimators & Basic Statistical Inference

Aleksandra Slavković

Department of Statistics
Penn State University

IPAM – DATA2010, Feb 25, 2010

Data privacy & Data analysis

Obtain valid statistical results while minimizing the loss of privacy and confidentiality of individuals and organizations.

Research Communities:

- Statistics: statistical disclosure limitation.
- Computer science: privacy-preserving data mining.

Nature of the problem has changed

Duality + Usability + Transparency

Data privacy & Data analysis

Obtain valid statistical results while minimizing the loss of privacy and confidentiality of individuals and organizations.

Research Communities:

- Statistics: statistical disclosure limitation.
- Computer science: privacy-preserving data mining.

Nature of the problem has changed
Duality + Usability + Transparency

Differential Privacy (DP) Framework

Precise guarantees on privacy in the presence of arbitrary side information, (possibly) in advance of data collection and publication.

Recent theoretical developments on connections between DP and traditional statistical inference

- Parametric estimation [Smith].
- Robust statistics [Dwork & Lei].
- Approximation of smooth densities [Wasserman & Zhou].

Our goals

- Understand how rigorous notions of privacy relate to statistical inference
- Evaluate how private and non-private estimators compare for parametric exponential families
- Evaluate the differential privacy framework to some popular statistical models such as log-linear models (contigency tables) or logistic regression models.
- Develop concrete methodology that data analysts can use

Clinical Trials

Clinial Trials:

- Data exchange: many confirmatory studies and careful meta-analyses are required to produce practical impact, i.e. changes to medical practice or public policy.
- Legacy: ClinicalTrials.gov is currently the largest registry in the world; it warehouses $86,148$ trials with locations in 172 countries as of today.
- Finite (typically small) sample size N.

Two research questions:

- How should we publish these current trial datasets for statistical analysis without compromising individual privacy?
- How should we design future trials to allow for such safe public sharing of results?

Outline

For this talk,

- focus on binomial distribution to evaluate the statistical efficiency of ML estimators and differential private estimators.
- illustrate the role of sample size in this interaction between statistical efficiency and privacy requirement.
- propose approximate sample size adjustment factors needed for sample size calculation in classical hypothesis testing.

Exponential Family

The exponential family density: $f(x|\theta) = h(x)exp\big(\sum_i \theta_i S_i(x) - K(\boldsymbol{\theta})\big)$

- $S_i(x)$s are sufficient statistics.
- θ_i's are natural parameters.
- $K(\boldsymbol{\theta})$ is the normalizing constant.

Consider a random sample $x_1, \ldots, x_N$ from $f(x|\theta)$. The Maximum Likelihood estimate of θ, $T_N(\mathbf{x})$, is obtained by maximizing the likelihood function $L(\theta) = \prod_{k=1}^{N} f(x_k|\theta)$.

$T_N(\mathbf{x})$ is a function of sufficient statistics $S_i(x)$s. Under the exponential family, all information of the random sample are contained in these sufficient statistics.

Asymptotic Efficiency of MLEs and Arbitrary Estimators

Theorem (Cramér) Let $X_1, X_2 \ldots$ be i.i.d with density $f(x|\theta), \theta \in \Theta$ and let θ_0 denote the true value of θ. Let the MLE of θ_0 be $T(x)$. Under appropriate regularity conditions:

$$\sqrt{N}(T_N(\mathbf{x}) - \theta) \xrightarrow{D} Normal(0, I^{-1}(\theta))$$

where $I(\theta)$ is Fisher information.

An arbitrary estimator $T_N^{\varepsilon}(\mathbf{x})$ is asymptotically efficient if it is also true that:

$$\sqrt{N}(T_N^{\varepsilon}(\mathbf{x}) - \theta) \xrightarrow{D} Normal(0, I^{-1}(\theta))$$

Mean Squared Errors

To compare the statistical quality of $T_N(\mathbf{x})$ and $T_N^{\varepsilon}(\mathbf{x})$ on a finite sample size, we can use the mean square error criterion:

$$
\begin{aligned}
MSE_{T_N(\mathbf{x})}(\theta) &= E_{\theta}\left[(T_N(\mathbf{x}) - \theta)^2\right] \\
MSE_{T_N^{\varepsilon}(\mathbf{x})}(\theta) &= E_{\theta}\left[(T_N^{\varepsilon}(\mathbf{x}) - \theta)^2\right]
\end{aligned}
$$

Data Access and Sharing

Communication between data servers and researchers:

- Datasets are contained in some centralized servers.
- Researchers access the servers to obtain sufficient statistics needed statistical inference.
- Differential privacy framework basically plays the role of a proxy by computing these statistics then adding Laplace noise to them before returning them to researchers.

Researchers can share data (or results) with others; e.g, in the context of clinical trials.

Focus on parametric inference with exponential families.

Neighboring Datasets & Differential Privacy

Two datasets $\mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{x'} = (x'_1, x'_2, ..., x'_n)$ are neighbors if and only if they are different at only one sample; i.e., rearrange $\mathbf{x} = (x_1, x_2, ..., x_i, ..., x_n)$ and $\mathbf{x'} = (x_1, x_2, ..., x'_i, ..., x_n)$ for some i in $1 \leq i \leq n$.

Definition

A statistic **T(.)** is ε-differentially private if for all neighboring datasets **x**, **x'**, and for all measurable subsets A:

$$\frac{P(T(\mathbf{x}) \in A)}{P(T(\mathbf{x'}) \in A)} \leq e^{\varepsilon}$$

The parameter $\varepsilon > 0$ is a measure of the information leakage.

Algorithm

Input: A data set $\mathbf{x} = (x_1, \ldots, x_N) \in D^N$.

Parameters:

- Λ is the range of $T_i(x)$, or diameter of the parameter space.
- $\varepsilon > 0$ is the level of privacy to achieve, i.e., perturbation parameter.

Algorithm 1:

- Obtain the sufficient statistics $T_1(x), \ldots, T_m(x)$.
- For each $T_i(x)$ draw a random observation R from $Laplace(\frac{\Lambda}{N\varepsilon})$ and compute $T_i^{\varepsilon}(x) = T_i(x) + R$.
- Return $T_i^{\varepsilon}(x)$'s.

$\varepsilon-$differential Privacy and Asymptotic Efficiency

[Smith] shows that privacy estimators theoretically achieve asymptotic efficiency when the sample sizes go to infinity.

Following lemmas are relevant for the binomial and multinomial models given our setting. We need to add more assumptions are needed for other models.

Lemma 1: Algorithm 1 satisfies $\varepsilon-$differental privacy.

Lemma 2: Under the regularity conditions of normal asymptotic distributions of ML estimators, if Λ is bounded and ε is fixed, the estimators $T_i^{\varepsilon}(x)$ are asymptotically unbiased, normal, and efficient.

The Triangle: MSE, ε, and N

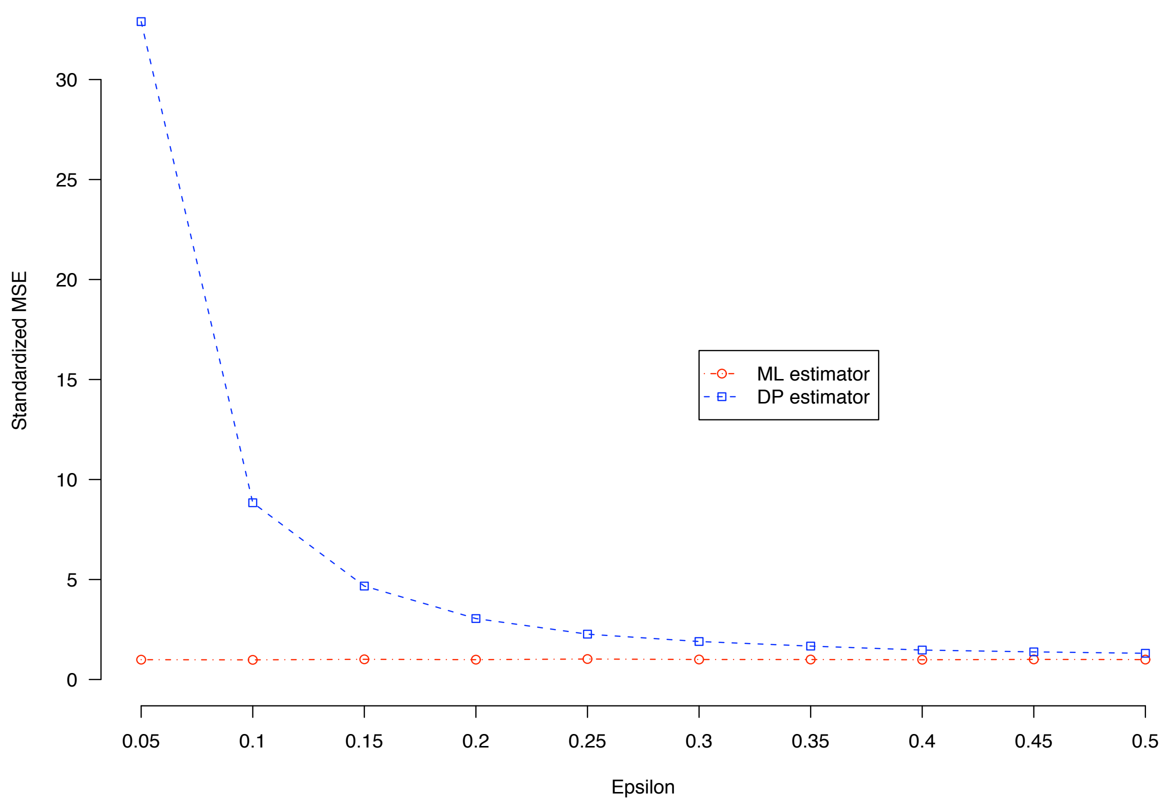
There are interactions among:

1. Quality of the estimator MSE.
2. Differential privacy parameter ε.
3. Sample size N.

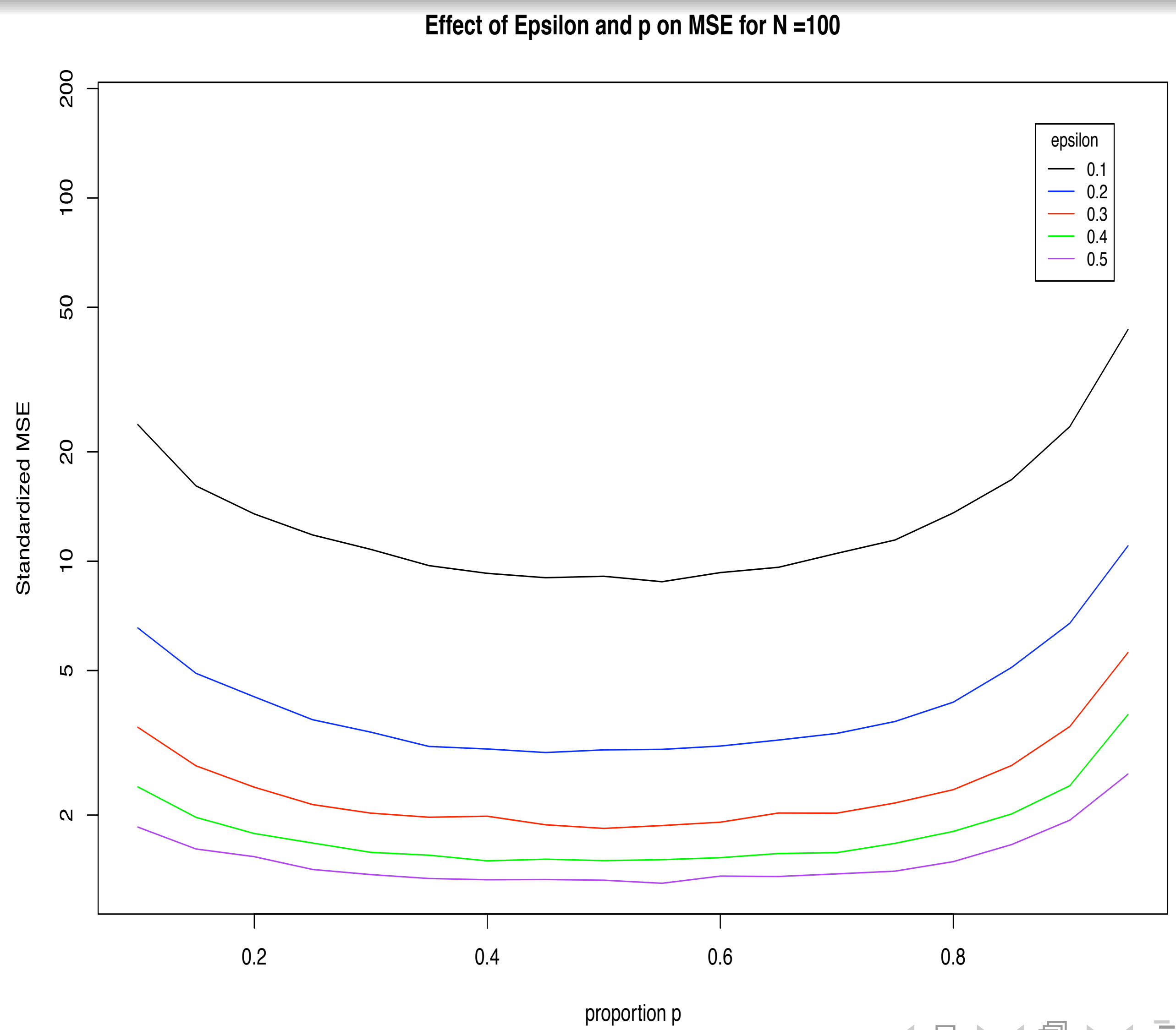
Since $T_N(\mathbf{x})$ is asymptotically unbiased, $MSE_{T_N(\mathbf{x})}(\theta) \approx Var\left[T_N(\mathbf{x})\right]$.
We will standardize both $MSE_{T_N(\mathbf{x})}(\theta)$ and $MSE_{T_N^{\varepsilon}(\mathbf{x})}(\theta)$ by $Var\left[T_N(\mathbf{x})\right]$.

Trade-off between Privacy and Efficiency through ε

Binomial: $p = 0.5$, sample size $N = 100$, simulation size $M = 10000$, $Lap(\frac{1}{N\varepsilon})$

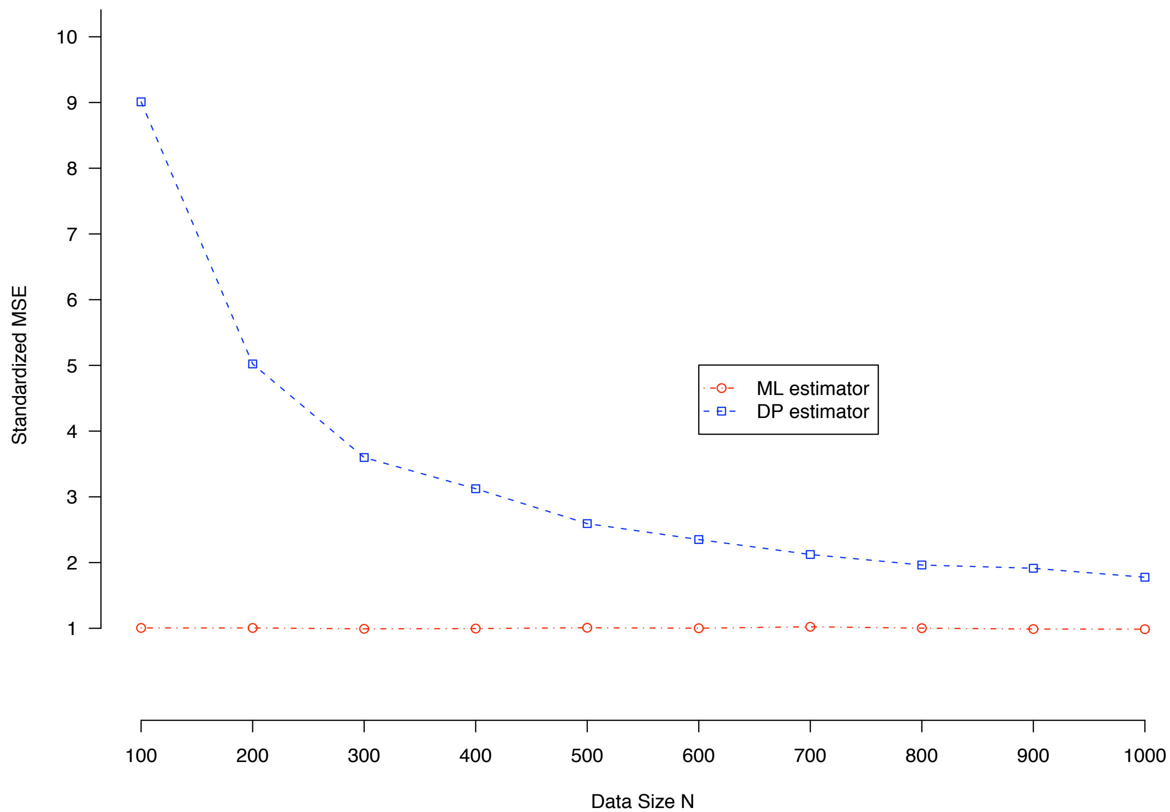


Trade-off between Privacy and Efficiency through ε

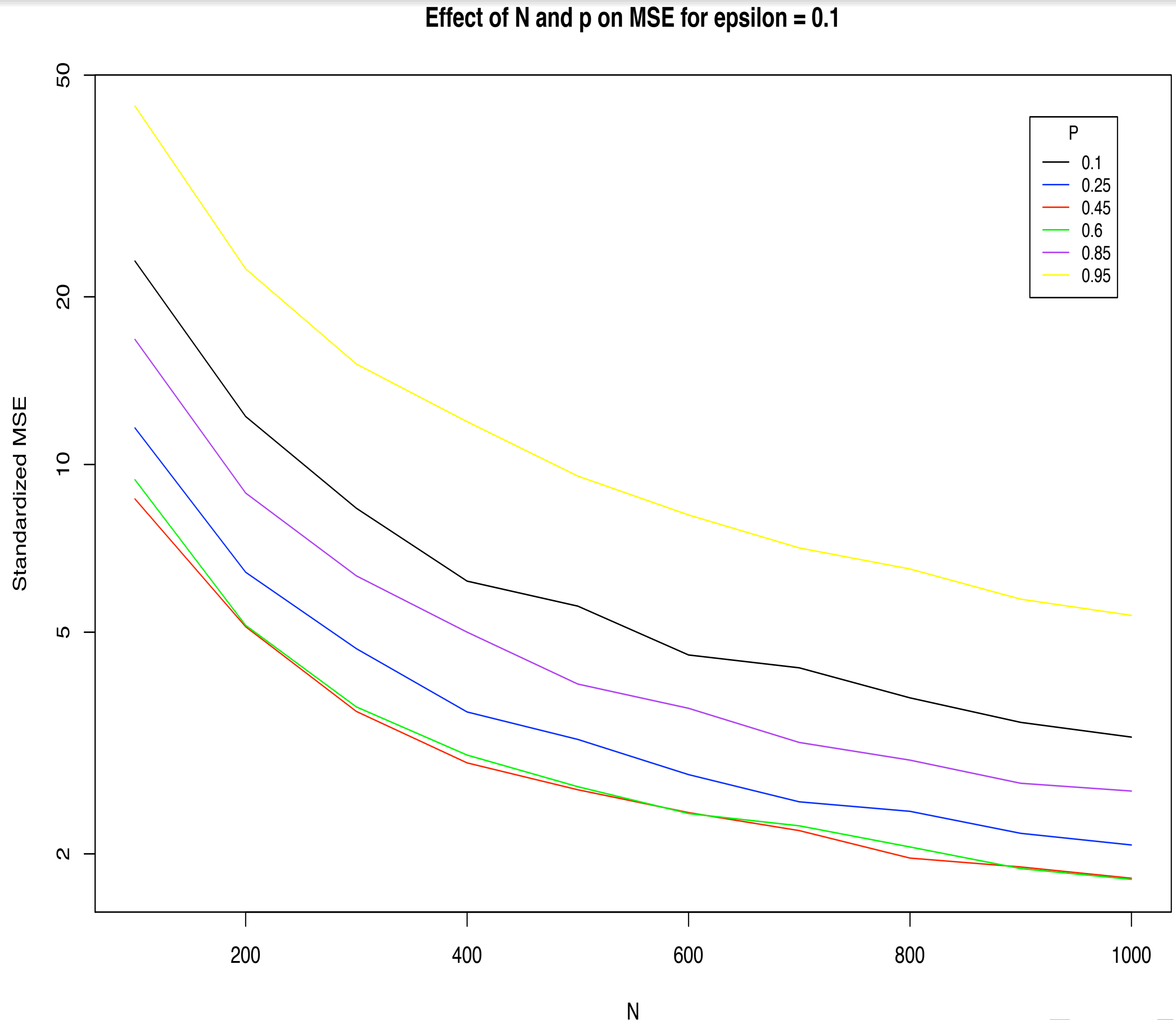


Achieving Asymptotic Efficiency by Controling Data Size

Binomial: $p = 0.5$, privacy level $\varepsilon = 0.1$, simulation size $M = 10000$, $Lap(\frac{1}{N\varepsilon})$



Achieving Asymptotic Efficiency by Controling Data Size



Concrete Methodology — Hypothesis Testing

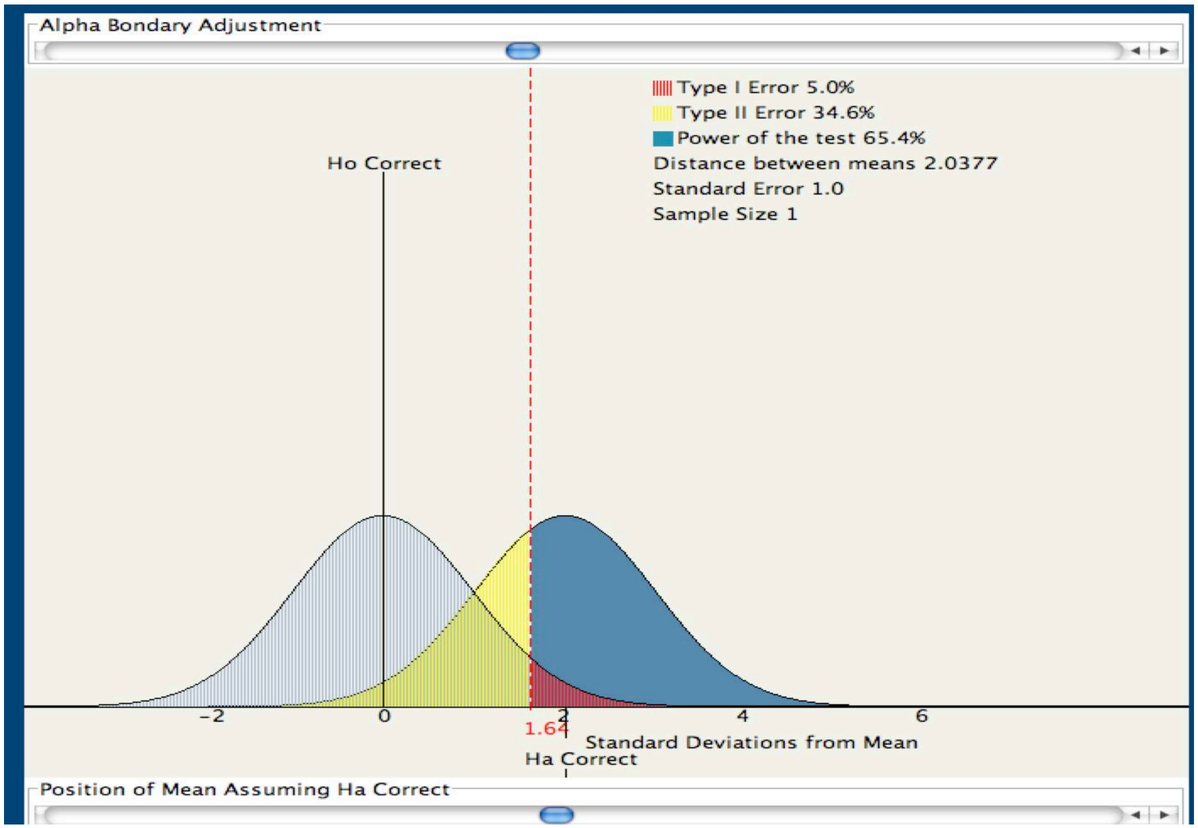
Many research questions in clinical trials are typically formulated in ways to test if we have sufficient evidence to reject some default theory or a null hypothesis in favor of a alternative hypothesis.

$$H_o : p = p_0 \text{ versus } H_a : p = p_0 + \delta.$$

Two criteria for comparing statistical hypothesis tests:

1. The confidence level of a test is defined as $1-\alpha$ where α, the type I error, is the probability of rejecting the null hypothesis when it is true.
2. The power of a test calculated as $1-\beta$ is the probablity of rejecting the null hypothesis when it is false (β is the type II error).

Concrete Methodology — Hypothesis testing



	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

$$\beta(\theta) = P_\theta(\mathbf{X} \in RR) \begin{cases} \text{Prob. of Type I error if } \theta \in \Theta_0 \\ \text{1- Prob. of Type II error if } \theta \in \Theta_0^c \end{cases}$$

Concrete Methodology – Hypothesis Testing

Many funding agencies and ethics boards frequently request a power analysis (sample size calculation) to be done before the study is conducted.

Two settings under the differential privacy:

- A priori determination of the revised finite sample size needed to achieve certain size and power of the test while maintaining the required differential privacy ε.
- If data are already available, researchers need to adjust original sample sizes for meta-analyses when calculation is based on differentially private sufficient statistics.

Test for a proportion

Let $x_1, x_2, \ldots, x_N \sim$ Bernoulli (p). The sufficient statistic $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is also the estimator of interest for p.

$$H_o : p = p_0 \text{ versus } H_a : p = p_0 + \delta.$$

N is the original sample size to achieve the confidence level $1-\alpha$ and the power $1-\beta$ in the case we do not deploy the differential privacy framework.

Under the differential privacy framework, to achieve the confidence level $1-\alpha$ and the power $1-\beta$, the privacy-preserving sample size $N^{\varepsilon} = K * N$.

Concrete Methodology – Hypothesis Testing

Simulations show that when the true data size is large enough the difference between N and N^{ε} is not significant.

We need to resolve the trade-off between statistical efficiency and privacy requirement by increasing the required sample size to control for the effect of noise.

Propose adjustment factors $K > 1$:

- A priori sample size determination: $N^{\varepsilon} = K * N$
- Meta-analyses: $N = N^{\varepsilon}/K$.

Without Differential Private Noise

Define $\sigma^2 = \bar{p}(1-\bar{p})$ where $\bar{p} = p_0 + \frac{\delta}{2}$

Under $H_o : \hat{p} \sim N\left(p_0, \frac{\sigma^2}{N}\right)$ versus Under $H_a : \hat{p} \sim N\left(p_0 + \delta, \frac{\sigma^2}{N}\right)$
To achieve the confidence level $1-\alpha$ and the power $1-\beta$, we solve:

$$p_0 + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{N}} = p_0 + \delta - z_{1-\beta}\sqrt{\frac{\sigma^2}{N}}$$

Then, the original sample size:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{\delta^2}$$

With Differential Privacy Noise

Under H_o :

$$N\left(p_0, \frac{\sigma^2}{N^\varepsilon}\right) + L\left(\frac{\sqrt{2}}{\varepsilon N^\varepsilon}\right) \approx N\left(p_0, \frac{\sigma^2}{N^\varepsilon} + \frac{2}{\varepsilon^2 (N^\varepsilon)^2}\right)$$

Under H_a:

$$N\left(p_0 + \delta, \frac{\sigma^2}{N^\varepsilon}\right) + L\left(\frac{\sqrt{2}}{\varepsilon N^\varepsilon}\right) \approx N\left(p_0 + \delta, \frac{\sigma^2}{N^\varepsilon} + \frac{2}{\varepsilon^2 (N^\varepsilon)^2}\right)$$

Here we approximate $L\left(\frac{\sqrt{2}}{\varepsilon N^\varepsilon}\right)$ by $N\left(0, \frac{2}{\varepsilon^2 (N^\varepsilon)^2}\right)$

Test for a proportion

The privacy-preserving sample size N^{ε} is calculated by solving:

$$p_0 + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{N^{\varepsilon}} + \frac{2}{\varepsilon^2(N^{\varepsilon})^2}} = p_0 + \delta - z_{1-\beta}\sqrt{\frac{\sigma^2}{N^{\varepsilon}} + \frac{2}{\varepsilon^2(N^{\varepsilon})^2}} \quad (1)$$

Then:

$$N^{\varepsilon} = N(\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{8\delta^2}{\varepsilon^2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^4}}), \quad (2)$$

Test for a proportion

Here we are interested in the approximate sample size correction factor under the DP framework:

$$K = \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{8\delta^2}{\varepsilon^2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^4}} \quad (3)$$

We can calculate a better approximate sample size correction factor K by solving the equation:

$$F_{X_o}^{-1}(1-\alpha/2) = F_{X_a}^{-1}(1-\beta) \quad (4)$$

with respect to the variable N^ε, where sampling distributions of $\hat{p}$ are $X_o \sim NL(p_0, \frac{\sigma^2}{N^\varepsilon}, \varepsilon N^\varepsilon, \varepsilon N^\varepsilon, 1)$, and $X_a \sim NL(p_0 + \delta, \frac{\sigma^2}{N^\varepsilon}, \varepsilon N^\varepsilon, \varepsilon N^\varepsilon, 1)$.

Sample size correction factor K

Table: $\alpha = .05$, $\beta = .4$, $p_0 = .25$, $\delta = .1$, classical sample size $N = 103$ and the DP sample size $N' = KN$.

ε	0.1	0.2	0.3	0.4	0.5
Norm-Lap K	3.65	2.12	1.64	1.42	1.29
Norm-Norm K	3.58	2.10	1.63	1.41	1.29

Table: $\alpha = .05$, $\beta = .1$, $p_0 = .25$, $\delta = .1$, classical sample size $N = 221$ and the DP sample size $N' = KN$.

ε	0.1	0.2	0.3	0.4	0.5
Norm-Lap K	2.62	1.64	1.35	1.22	1.15
Norm-Norm K	2.64	1.65	1.35	1.22	1.15

Sample size correction factor K

Table: $\alpha = .05$, $\beta = .4$, $p_0 = .25$, $\delta = .1$, classical sample size $N = 103$ and the DP sample size $N' = KN$.

ε	0.1	0.2	0.3	0.4	0.5
Norm-Lap K	3.65	2.12	1.64	1.42	1.29
Norm-Norm K	3.58	2.10	1.63	1.41	1.29

Table: $\alpha = .05$, $\beta = .1$, $p_0 = .25$, $\delta = .1$, classical sample size $N = 221$ and the DP sample size $N' = KN$.

ε	0.1	0.2	0.3	0.4	0.5
Norm-Lap K	2.62	1.64	1.35	1.22	1.15
Norm-Norm K	2.64	1.65	1.35	1.22	1.15

Sample size correction factor K

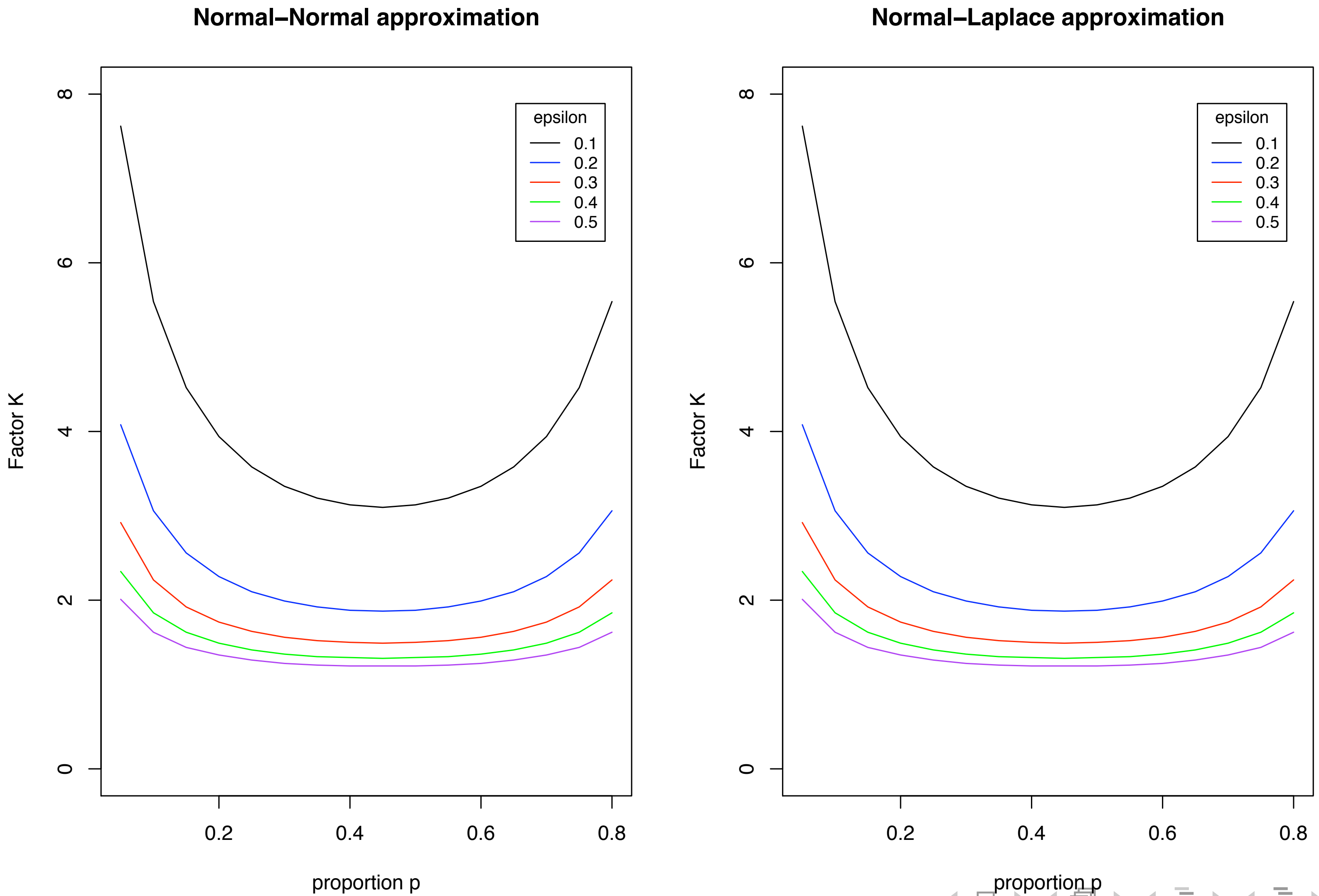
Table: The effect of correcting factors on achieving α with $\alpha = .05$, $\beta = .1$, $p_0 = .25$, $\delta = .1$. True sample size is $N = 221$.

ε	0.1	0.2	0.3	0.4	0.5
No Correction	0.1606	0.0763	0.0496	0.0350	0.0315
Norm-Appr K	0.0943	0.0557	0.0403	0.0376	0.0288
Norm-Lapl K	0.0937	0.0538	0.0387	0.0335	0.0267

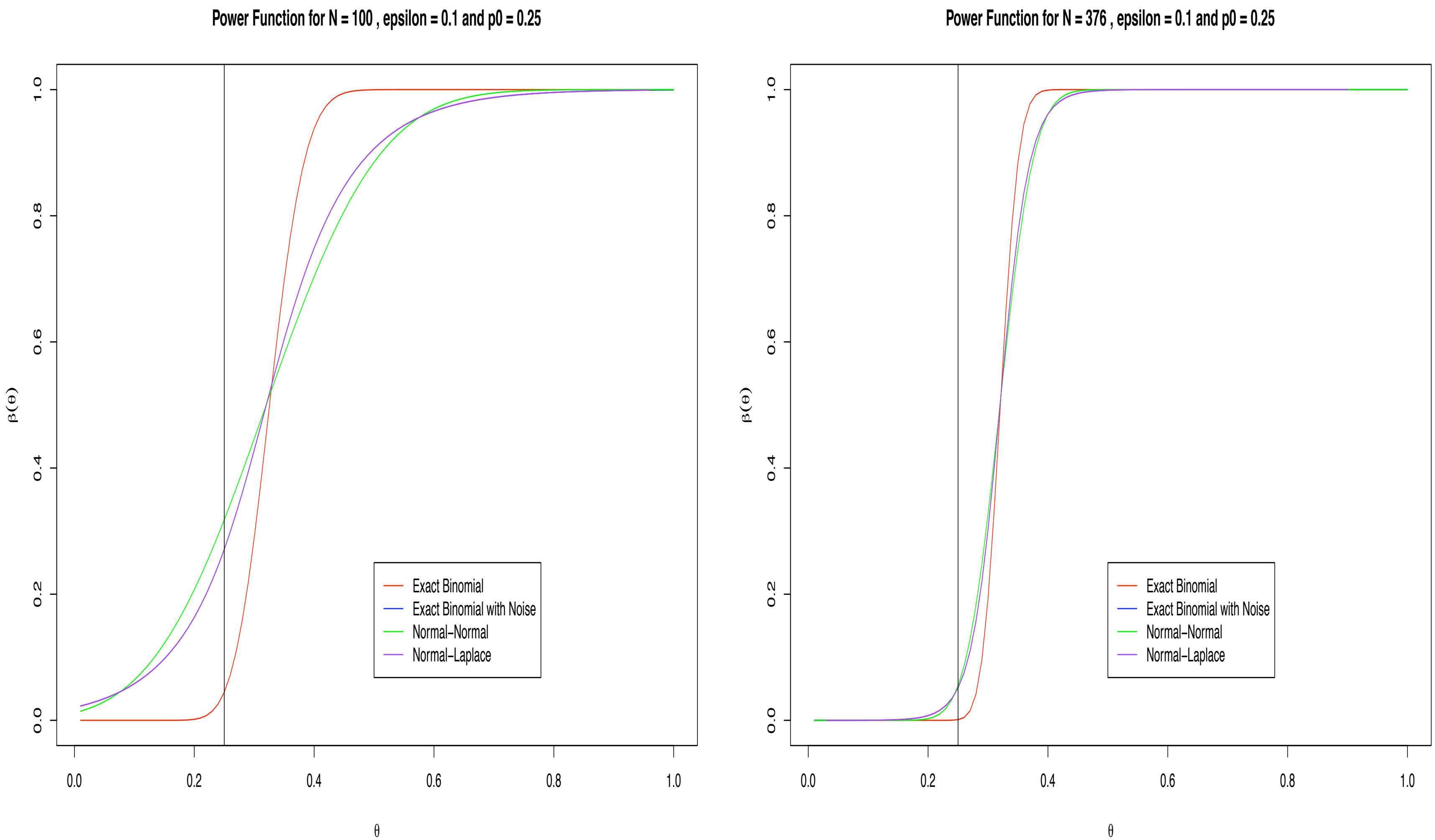
Table: The effect of correcting factors on achieving β with $\alpha = .05$, $\beta = .1$, $p_0 = .25$, $\delta = .1$. True sample size is $N = 221$.

ε	0.1	0.2	0.3	0.4	0.5
No Correction	0.2562	0.1817	0.1459	0.1355	0.1209
Norm-Appr K	0.0249	0.0434	0.0629	0.0754	0.0814
Norm-Lapl K	0.0238	0.0451	0.0633	0.0752	0.0830

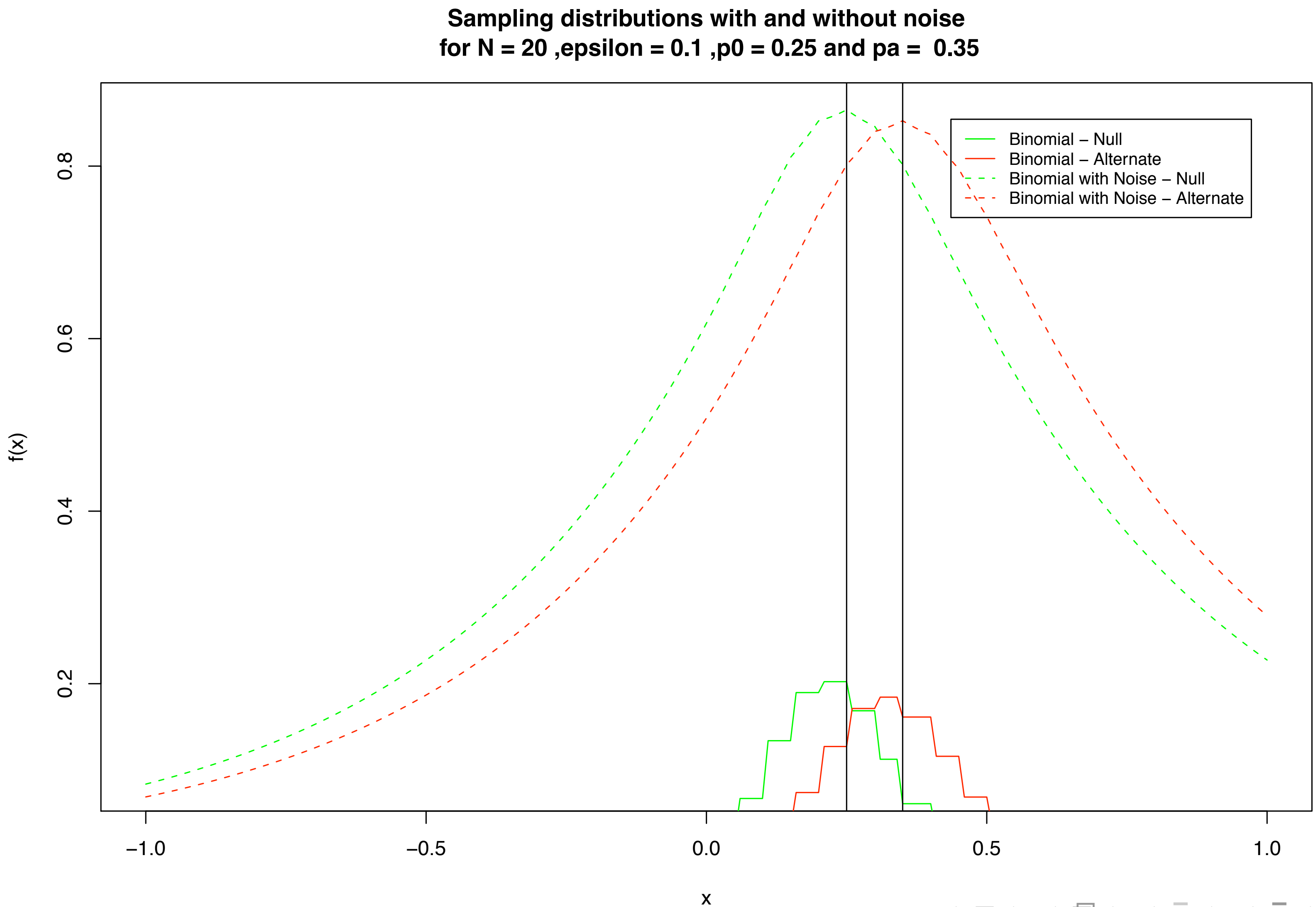
Sample size correction factor K



Power functions



Exact Methods for small N



Conclusions

Current results:

- Evaluate the effect of the data size on the asymptotic efficency of $\varepsilon-$differential estimators for Binomial and Multinomial parameters.
- Develop rules for sample size calculation and power analysis
  - Frequentist testing for a single proportion
  - χ^2 test of independence

Ongoing and future work:

- Evaluate the $\varepsilon-$differential privacy framework for log-linear models and logistic regression models.
- Apply differential privacy to Bayesian credible intervals.
- New statistical tests

Thank you!

Acknowledgments:

- Duy Vu and Vishesh Karwa
- Adam Smith
- NSF SES-0532407 and ONR-MURI-N00014-08-1-1015 to the Department of Statistics, Penn State University

References:

- Vu, D & Slavkovic, A. (2009) *Differential Privacy for Clinical Trial Data: Preliminary Evaluations*
- Smith, A. (2008) *Efficient, differentially private point estimators.*
- Dwork, C. & Lei, J. (2009) *Differential Privacy and Robust Statistics.*
- Wasserman, L. & Zhou, S. (2008) *A statistical framework for differential privacy*.