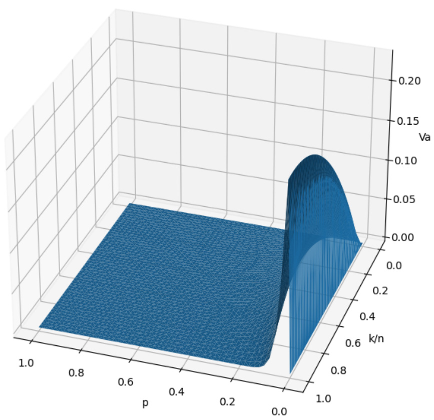
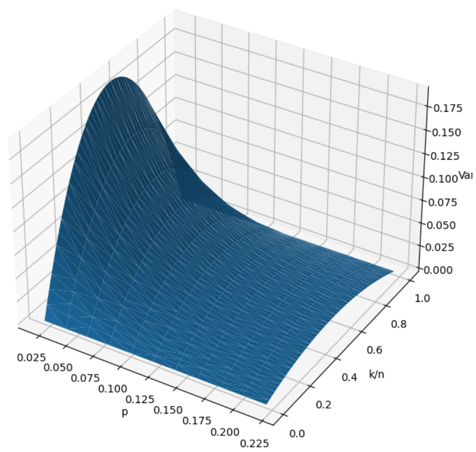
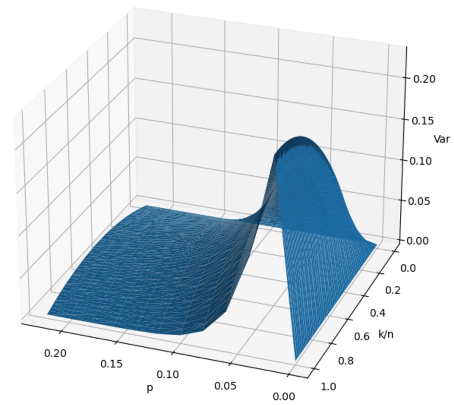
Variance of Estimator (See proof section for how plots were generated)

A plot of the variance of the estimator as a function of p=m/n and k/n.

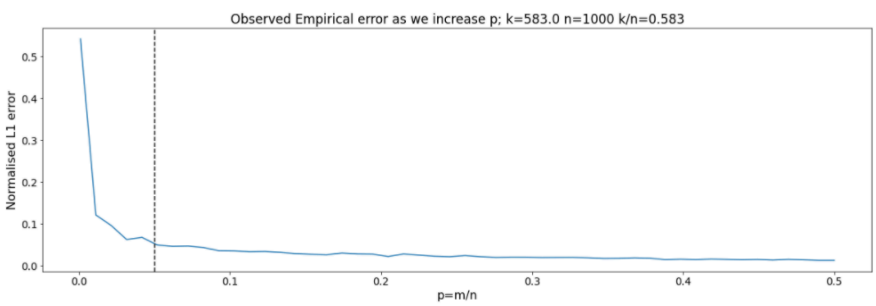




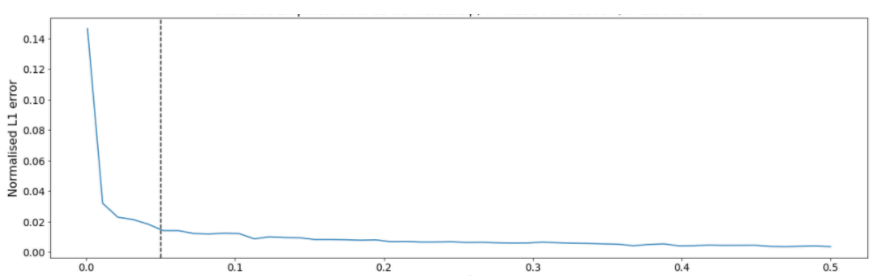


As p increases from 0 to 1, there seems to a cutoff point -- lets call it p_hat after which the variance drops to near 0 and is remains quite flat. In the words of the sample and threshold algorithm, this is saying that for there to be any utility, m has to be greater than a certain value. However, after that, it's really diminishing returns. In other words, I once I've seen enough of a sample, seeing more really does not offer me that much benefit. The above theoretical property can be empirically validated. For a given p, we ran the sample and threshold algorithm 100 times and plot the average error as p increases from 0 to 0.5. In the figure below we fixed n=1,000 and k/n = 0.583

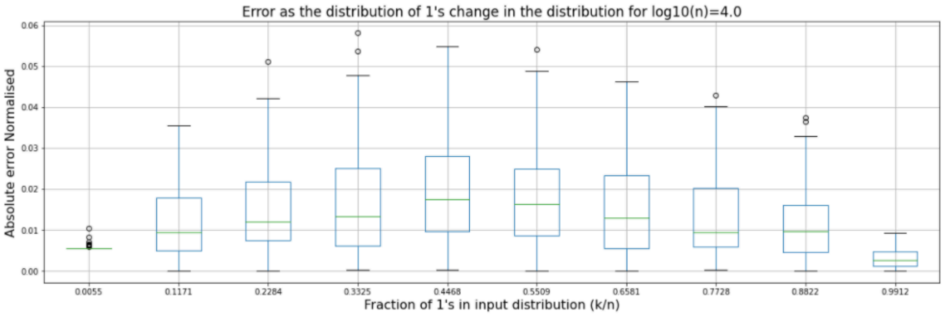
EXP I



EXP II: n=10,000, k/n = 0.298

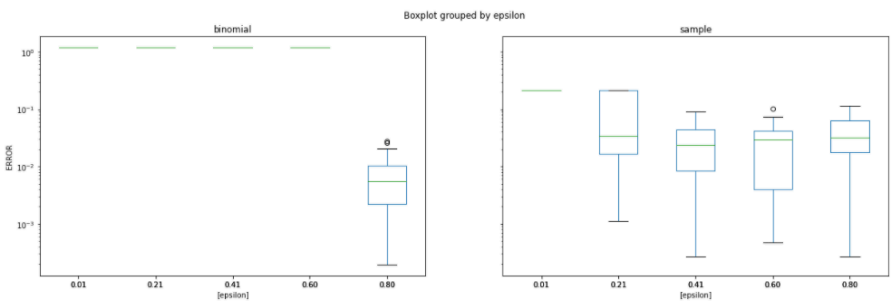


Exp III: p=0.05, n=10,000. We plot the empirical error as we vary the number of 1s in a dataset. The general trend of the curve is as we see in the variance plots.



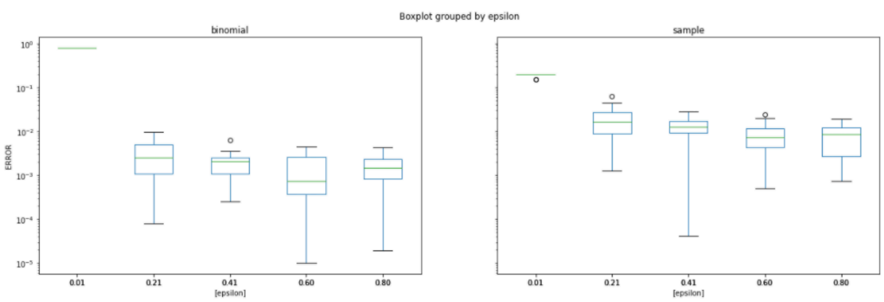
Comparison of Sample and Threshold vs Binomial noise

However a disadvantage is that as n increases, my error (variance) does not really shrink to 0 as fast as it does for binomial/subgaussian noise. More imporantly, when n is large, the error is almost is worse than that of adding binomial noise. The error of the binomial noise procedure is independent of k/n. The variance of estimator for binomial noise goes to 0 as n goes to infinity. Big n does not inhibit privacy. In contrast the error of sample and threshold goes to 0 as p goes to 1 -- but that prevents privacy.



n = 1000
k/n = 0.1988

For medium sized datasets, we cannot even use the binomial noise distribution.



n = 10,000
k/n = 0.1988

Proof For Variance

We have $x_1, \dots x_n \in \{0,1\}$ where $\sum_{i=1}^{n} x_i = k$.

We also have $z_i \sim \text{Bernoulli}(p)$. We want to estimate k/n. Our estimator is the following:

$$\tilde{x} = \frac{\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} z_i} = \frac{\#\text{ of 1's in sampled list}}{\#\text{ of items sampled.}}$$

Let $f(a) = \mathbb{P}(\tilde{x} = a)$:= Probability mass function.

$\boxed{\text{Note}}$: if $\sum_{i=1}^{n} z_i = 0$ then we re-run the experiment of Bernoulli sampling over n items.

$$f(a) = \sum_{s=1}^{n} \mathbb{P}\left[\frac{\sum_{i=1}^{n} x_i z_i}{s} = a \,\middle|\, \sum_{i=1}^{n} z_i = s\right] \mathbb{P}(s)$$

$$+ (1-p)^n f(a).$$

The second summand is just the probability of all z_i's being 0.

$$P(s) = \binom{n}{s} p^s (1-p)^{n-s}$$

↳ just the binomial distribution.

Let $a = \frac{b}{s}$; then we have

$$f(a)\left[1-(1-p)^n\right] = \sum_{s=1}^{n} \mathbb{P}\left[\frac{\sum_{i=1}^{n} x_i z_i}{s} = \frac{b}{s} \,\Big|\, \sum z_i = s\right] P(s)$$

$$f(a) = \frac{1}{\left[1-(1-p)^n\right]} \sum_{s=1}^{n} \mathbb{P}\left[\sum x_i z_i = b \,\Big|\, \sum z_i = s\right] P(s)$$

$$\mathbb{P}\left[\sum x_i z_i = b \,\Big|\, \sum z_i = s\right] = \frac{\binom{k}{b}\binom{n-k}{s-b}}{\binom{n}{s}}$$

PMF :=

$$\boxed{\therefore\ f(a) = \frac{\sum_{s=1}^{n} \binom{k}{b}\binom{n-k}{s-b} p^s (1-p)^{n-s}}{\left[1-(1-p)^n\right]}}$$

$$\boxed{Var(\tilde{X}) = E[\tilde{X}^2] - E[X]^2}$$

$$\boxed{E[\tilde{X}] = \sum_{s=1}^{n} \sum_{b=0}^{s} \frac{b}{s} f(b/s)}$$

↑ plug into computer and plot !