Bias Reduction for Sum Estimation

Will Rosenbaum (Amherst College)
Joint with: Talya Eden, Jakob Hauen, Shyam Narayanan, and Jakub Tetek
Overview

1. Motivation
2. General Setting
3. Bias Reducing Sum Estimation
4. Lower Bounds
Motivation
Motivating Example

Population of $N$ voters deciding a proposition (yes/no)

Goal. Estimate the number of “yes” votes.

- expected margin of victory is $\sim 3\%$

Typical Assumption (unfounded?). We can sample voters uniformly at random.
Easy Estimation

Procedure:

• sample $m$ voters
• $X_i \in \{0, 1\}$ is vote of $i$th sample
• return

\[
\bar{\mu} = \frac{1}{m} \sum_{i=1}^{m} N X_i.
\]

Guarantee:

• $\varepsilon N$ additive error with probability $2/3$ for $m = O(\varepsilon^{-2})$. 

Estimating Yes Votes

\[ N = 108, \ m = 20, \ \mu = 58 \]

- \( \bar{\mu} = \frac{1}{m} \sum_i N X_i = 59.4 \)
- “Yes” predicted to win!
More Realistic Assumption?

Samples are not *exactly* uniform

- there may be some underlying bias
  - bias may systematically favor “yes” or “no” votes
- true distribution is *not known*...
More Realistic Assumption?

Samples are not *exactly* uniform

- there may be some underlying bias
  - bias may systematically favor “yes” or “no” votes
- true distribution is *not known*...
- ...but true distribution is *close* to uniform
  - e.g., each voter sampled with probability \((1 \pm 0.1)/N\)
  - *bias* is \(0.1 = 10\%\)
“Yes” Bias

$N = 108, m = 20, \mu = 58$

“Yes” voters 10% more likely to be sampled

- $\bar{\mu} = \frac{1}{m} \sum_i N X_i = 64.8$
- “Yes” predicted to win!
“No” Bias

\[ N = 108, \ m = 20, \ \mu = 58 \]

“No” voters 10% more likely to be sampled

- \[ \bar{\mu} = \frac{1}{m} \sum_i N X_i = 48.6 \]
- “Yes” predicted to lose!
Can We Do Better?

**Question.** Can we estimate $\mu = \sum_i x_i$ to within additive error, say, $0.02N$ using a sub-linear number of samples?

- naive estimates would have error proportional to the sample bias
Yes We Can!

Our Results (special case):

1. If bias is 0.1, 0.02N additive error is achievable using $O(\sqrt{N})$ samples

2. More generally, estimating $\sum x_i$ with $x_i \in \{0, 1\}$:
   - $\gamma$ is upper bound on sample bias
   - $\varepsilon$ is desired approximation factor
   - $k = \lceil \log(\varepsilon)/\log(\gamma) \rceil$
   - $O(N^{1-1/k})$ samples are sufficient

3. This sample complexity is tight
Bigger Picture

In many contexts, it is assumed that samples can be generated *exactly* from some distribution:

- statistics
- sub-linear time algorithms
- distribution testing

General Questions

1. Under what circumstances are “noisy” samples sufficient?
2. What is the algorithmic cost of coping with noise?
General Setting
Setup

- universe of $N$ elements, $[N] = \{1, 2, \ldots, N\}$, item $i$ has associated value $x_i$
- $\mathcal{P}$ a known probability distribution over $[N]$
- $\mathcal{Q}$ is true sample distribution
- $\mathcal{Q}$ is pointwise $\gamma$-close to $\mathcal{P}$:

\[
(1 - \gamma) \mathcal{P}(i) \leq \mathcal{Q}(i) \leq (1 + \gamma) \mathcal{P}(i) \quad \forall i
\]

Goal. Estimate $\mu = \sum_i x_i$ to within additive error $\varepsilon_1 \mu_+ + \varepsilon_2$, where

- $\mu_+ = \sum_i |x_i|$
- $\varepsilon_1 \ll \gamma$
Classical Setting \( (\mathcal{P} = \mathcal{Q}) \)

**Hansen-Hurwitz Estimator**

Sample \( m \) elements with replacement

- unbiased estimator when \( \mathcal{P} = \mathcal{Q} \)
- bias is at most \( \gamma \mu_+ \) when \( \mathcal{Q} \) is \( \gamma \)-close to \( \mathcal{P} \)

\[
\mu_{HH} = \frac{1}{m} \sum_{j=1}^{m} \mathcal{P}(i_j)^{-1} X_j
\]
Bias-Reducing Sum Estimation
Our Idea

Apply Hansen-Hurwitz to \textit{h-wise collision statistics}

- \textit{h-wise collision} is a set of \textit{h} equal sampled indices
  \[ i_1 = i_2 = \cdots = i_h \]

Get \textit{h-wise collision} estimators:

- \[
  \xi_h = \frac{1}{\binom{m}{h}} \sum_{i=1}^{N} \left( \begin{array}{c} Y_i \\ h \end{array} \right) (P(i))^{-h} x_i
\]

- \( Y_i = \# \) of times index \( i \) was sampled

Note. \( \xi_1 = \mu_{HH} \).
Letdown

$h$-wise collision estimators alone are no better than Hansen-Hurwitz...

- bias is $O(\gamma h \mu_+)$
- variance is also worse
Letdown

$h$-wise collision estimators alone are no better than Hansen-Hurwitz...

- bias is $O(\gamma h \mu_+)$
- variance is also worse

...but an appropriate linear combination of these estimators result in lower-order errors (in $\gamma$) to cancel out...
Bias Reducing Estimator

We define the **bias reducing estimator** $\zeta_h$:
- $\zeta_k = \sum_{h=1}^{k} (-1)^{h+1} \binom{k}{h} \xi_h$

**Main Result.** $\zeta_k$ has bias at most $\gamma^k \mu_+$. 
- we also bound the variance of $\zeta_k$. 
A Consequence

Special Case: $\mathcal{P}$ is uniform, all values $x_i \in \{0, 1\}$

- take

$$m = O\left(\sqrt{n^{k-1} \varepsilon_{2}^{-2} \text{Var}_{HH}}\right)$$

- then $m$ samples are sufficient to estimate $\mu$ to additive error $\gamma^k \mu + \varepsilon_2$ with probability $2/3$

In original example

- $\gamma = 0.1$,
- $\varepsilon_1 = 0.01 = \gamma^2$
- $\varepsilon_2 = 0.01N$

$\implies m = O(\sqrt{N})$ samples are sufficient
Lower Bounds
Setting of Original Example

- \( \mathcal{P} = \) uniform
- \( x_i \in \{0, 1\} \)

Main Result (lower bound). Suppose algorithm \( A \) guarantees:

- for any desired \( \varepsilon > 0 \),
- for every distribution \( Q \) that is point-wise \( \gamma \)-close to uniform
- \( A \) returns an \( \varepsilon N \) additive approximation to \( \mu \) with constant probability

Then: for every positive integer \( k \), there exists a constant \( c_k < 1 \) such that for \( \varepsilon \leq c_k \gamma^k \), \( A \) requires \( \Omega(N^{1-1/(k+1)}) \) samples.
Bounds in Pictures
Questions

1. Tighten values of $c_k$

2. Sample correction: can we reduce bias of individual samples? Can we do this in an amortized fashion?

3. Can other algorithms/estimation tasks be made robust to sample bias?
Thank You!!!

Any questions?
Appendix 1: Estimator for $P = \text{uniform}, k = 2$
Bias Reducing Estimator, $k = 2$

General case:

- $\zeta_k = \sum_{h=1}^{k} (-1)^{h+1} \binom{k}{h} \xi_h$

Case $k = 2$:

- $\zeta_2 = 2\xi_1 - \xi_2$

Suppose $Q(i) = (1 + \alpha_i)/N$, $|\alpha_i| \leq \gamma$

Then:

- $E(\xi_1) = \sum_i (1 + \alpha_i)x_i$
- $E(\xi_2) = \sum_i (1 + \alpha_i)^2x_i$

So:

- $E(2\xi_1 - \xi_2) = \sum_i (2(1 + \alpha_i) - (1 + \alpha_i)^2)x_i = \mu + \sum_i \alpha_i^2 x_i$
Appendix 2: Lower Bound Techniques
Step 1: Distinguishing Distributions

**Theorem** (Raskhodnikova et al. 2009). If two distribution \( D_1 \) and \( D_2 \) have same \( p \)th frequency moments for \( p = 1, 2, \ldots, k \), then \( \Omega(N^{1-1/(k+1)}) \) samples are required to distinguish \( D_1 \) from \( D_2 \).

**We Construct.** Distributions \( D_1 \) and \( D_2 \) with:

- support size \( n_1 = (1 + O(\gamma^k))n \) and \( n_2 = n \) (respectively)
- both pointwise \( \gamma \)-close to uniform
- identical frequency moments \( p = 1, 2, \ldots, k \)
Step 2: Sum Estimation

Given $D_1$ and $D_2$, set $N = n_1 + n_2$, $Q = \frac{1}{2} D_1 + \frac{1}{2} D_2$

- Scenario 1:
  
  $$x_i = \begin{cases} 
  1 & i \in \text{supp}(D_1) \\
  0 & i \in \text{supp}(D_2) 
  \end{cases}$$

- Scenario 2:
  
  $$x_i = \begin{cases} 
  0 & i \in \text{supp}(D_1) \\
  1 & i \in \text{supp}(D_2) 
  \end{cases}$$

Then estimating $\sum_i x_i$ with error $O(\gamma^k)$

$\implies$ distinguishing scenarios 1 and 2

$\implies$ distinguishing $D_1$ and $D_2$