Prior-free Data Acquisition for Accurate Statistical Estimation

Yiling Chen, **Shuran Zheng**

Harvard University

June, 2019

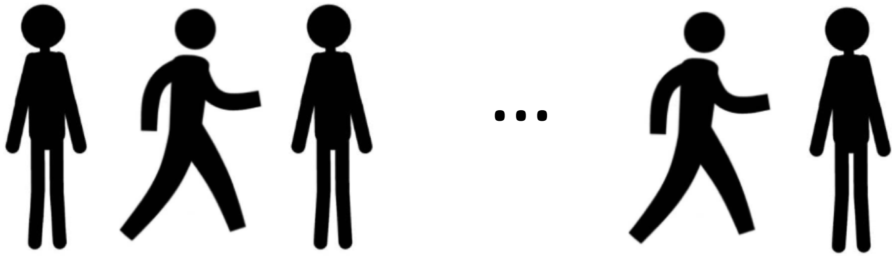
Acquiring data from self-interested individuals to estimate some statistic of a population

Problem description

A data analyst

- Avg. daily workout time?
- Budget B
- No prior information about the cost or data

n data providers



- Incur cost to record workout time
- Cost and data arbitrarily correlated

Model

A data analyst

- Estimate the mean of some parameter of interest z
- Budget B
- No prior information of the cost (or data)

n data providers

- Incur cost c_i to acquire the data z_i
- Cost and data arbitrarily correlated
- Self-interested

- For $i = 1, \ldots, n$
  1. The i-th data provider arrives (in random order).
  2. Decide a mechanism M_i to purchase the i-th data point z_i based on all observed history H_{i-1} ,
- Aggregate all collected information to output an estimator S of the population mean $\frac{1}{n}\sum z_i$.

Problem description

- For $i = 1, \ldots, n$
  1. The i-th data holder arrives (in random order).
  2. Based on all observed history H_{i-1} , decide a mechanism M_i to purchase the i-th data point z_i.
- Aggregate all collected information to output an estimator S of the population mean $\frac{1}{n}\sum z_i$.

- Objective: output a good estimator S
  - Unbiased point estimation: small variance
  - Interval estimation: minimize the length
- Constraint: expected spending $\leq$ budget B

Previous results: known cost distribution

A simpler problem *Roth and Schoenebeck [2012], Chen et al. [2018]:*

- the marginal cost distribution is known
- find a fixed mechanism to purchase n data points
- unbiased estimator with minimum variance (in worst-case cost-data correlation)

Previous results: known cost distribution

Naïve purchasing mechanisms:

- Fixed price p so that the expected spending $= B$

Bias toward the low cost sub-population!

- Purchase with a constant probability q, output $\sum \widehat{z_i}/q$

Variance may not be optimal

Survey mechanisms from Roth and Schoenebeck [2012]

- Purchase data with different costs with different probabilities and prices

(c_i, z_i)

report $\hat{c}_i$

Purchase with probability $A(\hat{c}_i) > 0$

Pay $P(\hat{c}_i) \geq \hat{c}_i$ and observe $\hat{z}_i = z_i$

Not purchase with probability $1 - A(\hat{c}_i)$

$\hat{z}_i = 0$

Horvitz-Thompson estimator: $\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{z}_i}{A(\hat{c}_i)}$

Previous results: known cost distribution

Horvitz-Thompson estimator: $\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{z}_i}{A(\hat{c}_i)}$

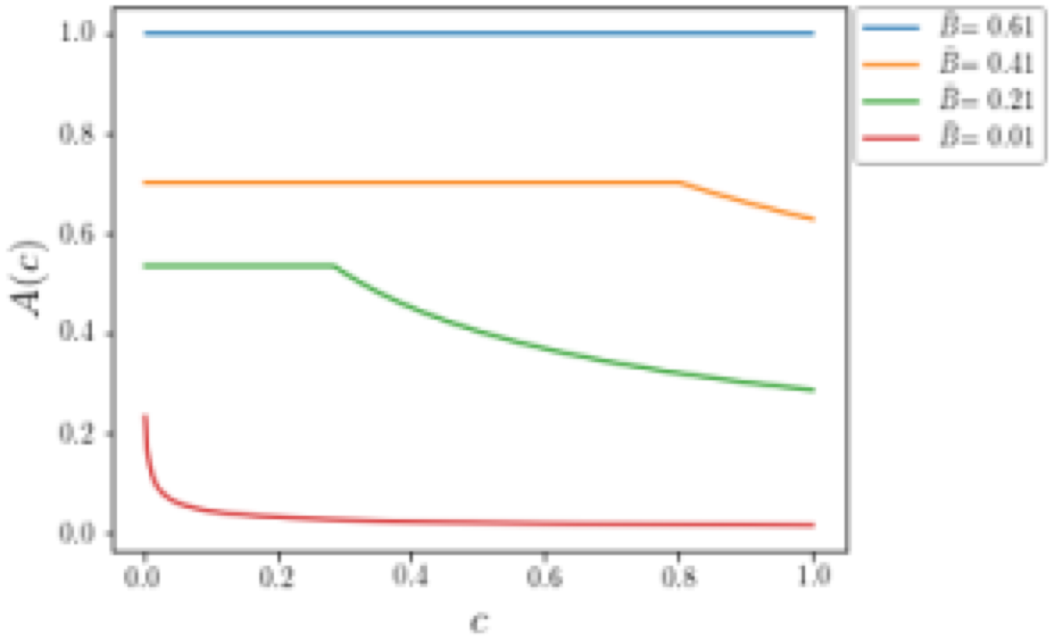
- Minimize the variance of the output Horvitz-Thompson estimator.
- $A(c)$, $P(c)$ should satisfy
  - Individual rationality: $P(c) \geq c$
  - Incentive compatibility
  - Budget feasibility: $E_c[A(c)P(c)] \leq \overline{B}$

$OPT(n, C, B) = (A^*, P^*)$

$C \;= \{c_1, \ldots, c_n\}$

Previous results: known cost distribution

- Characterization of $A^*(c)$ from Chen et al. [2018]
- Virtual costs $\phi(c)$
- $A^*(c) \propto \frac{1}{\sqrt{\phi(c)}}$



Unknown cost distribution: challenges

- $OPT(n,$ ~~C~~$, B) =$ ~~(A^*, P^*)~~
- Make purchasing decisions **without knowing future costs**
  - satisfy the budget constraint
  - optimize the performance
- Adjust the mechanism based on the observed costs

Our contribution

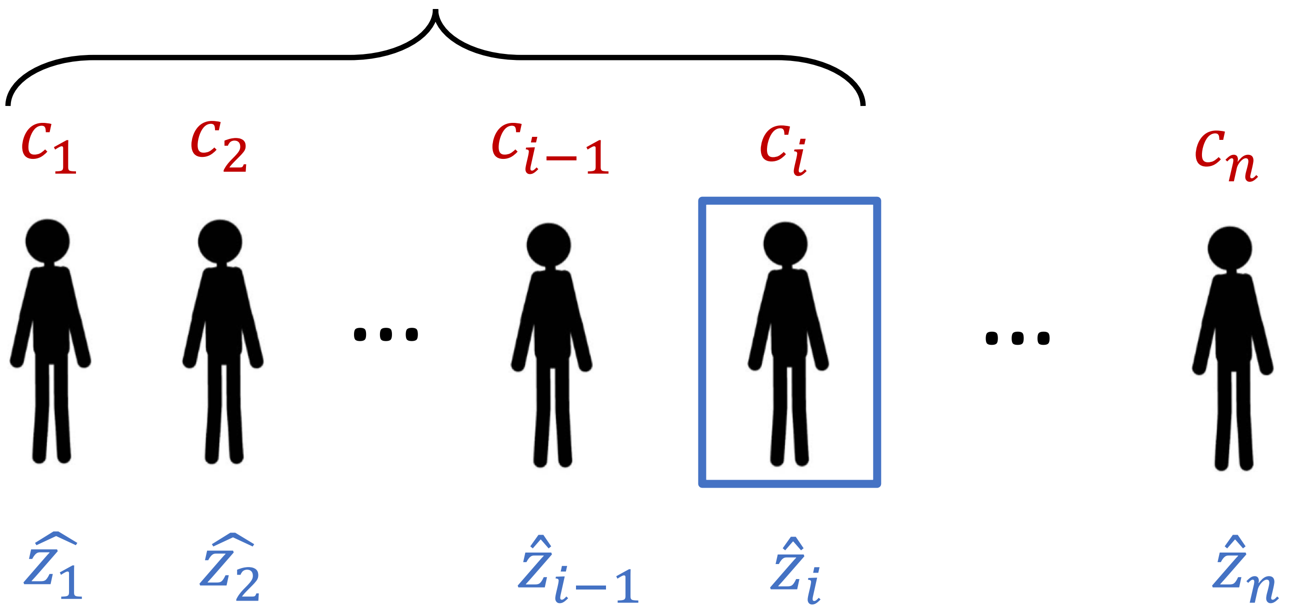
- Prior-free mechanism design
  - Performance matches that of the optimal mechanism, which knows the true cost distribution, within a constant factor.

- Confidence interval estimator

Prior-free mechanisms: algorithm

- At round i, use a survey mechanism M_i

$$APPROX\left(i, \{c_1, \ldots, c_{i-1}, \overline{C}\}, B_i\right) \approx OPT\left(i, \{c_1, \ldots, c_{i-1}, \overline{C}\}, B_i\right)$$



Output $\frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{z}_i}{A^i(\widehat{c}_i)}$

Prior-free mechanisms: result

Theorem: When we use $B_i \propto \sqrt{i}$, our mechanism is

- IC and IR,
- with expected total spending no more than B,
- and performance no worse than a constant factor times the benchmark $OPT(n+1, \{c_1, \ldots, c_n, \overline{C}\}, B)$.

Prior-free mechanisms: proof ideas

- At round i, use a survey mechanism M_i

Step #1: Decompose the variance into per-round ``loss''

Variance of $S \approx$ $\mathrm{E}[\text{loss}(M_1)] + \mathrm{E}[\text{loss}(M_2)] + \cdots + \mathrm{E}[\text{loss}(M_n)]$

$$\mathrm{E}[\text{loss}(M_i)] = \mathrm{E}\left[\frac{1}{A^i(\hat{c}_i)}\right]$$

Prior-free mechanisms: proof ideas

Step #2: Compare the loss of our mechanism with the loss of the benchmark

- $L(n, C, B) =$ expected loss of using $APPROX(n, C, B)$ when the data holder's cost is randomly chosen from C

Properties of $L(n, C, B)$:

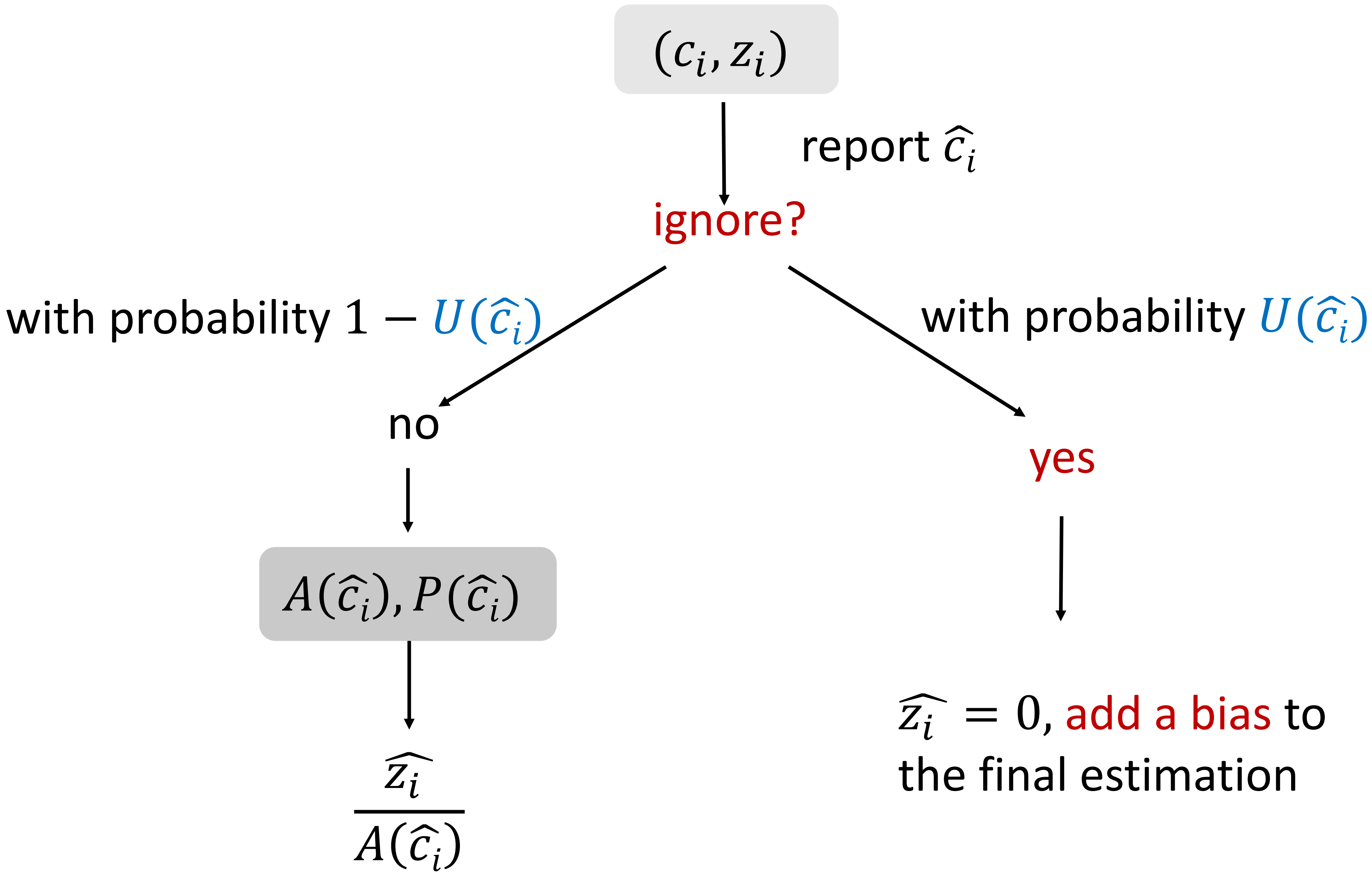
1. $L(n, C, B/k) \leq k \cdot L(n, C, B)$ for any n, C, B, k
2. Let S be a random subset $\subseteq C$ with $|S| = k$,

$$E_S[L(k, S, B)] \leq L(n, C, B)$$

- At round i, allocate budget B_i, the expected loss $\leq L(n, C, B) \cdot \frac{B}{B_i}$
- Choose $B_i \propto \sqrt{i}$, total "loss" $\leq$ constant * benchmark

Confidence interval estimator

- Allow the estimator to be biased
- Ignore some high-cost data points
- Bias-variance tradeoff
- Optimal confidence interval: minimize the worst-case expected length.



(c_i, z_i)
report $\widehat{c_i}$
ignore?
with probability $1 - U(\widehat{c_i})$
with probability $U(\widehat{c_i})$
no
yes
$A(\widehat{c_i}), P(\widehat{c_i})$
$\frac{\widehat{z_i}}{A(\widehat{c_i})}$
$\widehat{z_i} = 0$, add a bias to the final estimation

Confidence interval estimator

- Characterization of a 2-approximation of the optimal confidence interval when the cost distribution is known
- Online mechanism that matches the benchmark within a constant factor

Thanks & Questions?

First estimate the costs

- Truthfulness guarantee weaker
- Difficult to estimate $\phi(c)$

Questions

- Bandits with knapsack (dynamic pricing):
  - Action space too large
  - Regret dependent on |A|

- Online convex optimization
  - Put the violation of budget constraint into objective function: cannot be decomposed into per round loss function
  - Online convex optimization (with long-term budget): unknown budget constraint-> unknown X