CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 17

Samson Zhou

Previously: Variance

- The variance of a random variable X over Ω is:

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

- Linearity of variance for *independent* random variables: $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Previously: Chebyshev's Inequality

- Let X be a random variable with expected value $\mu \coloneqq \mathrm{E}[X]$ and variance $\sigma^2 \coloneqq \mathrm{Var}[X]$

- $\Pr[|X - \mathrm{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$ becomes $\Pr[|X - \mathrm{E}[X]| \geq t] \leq \frac{\sigma^2}{t^2}$

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

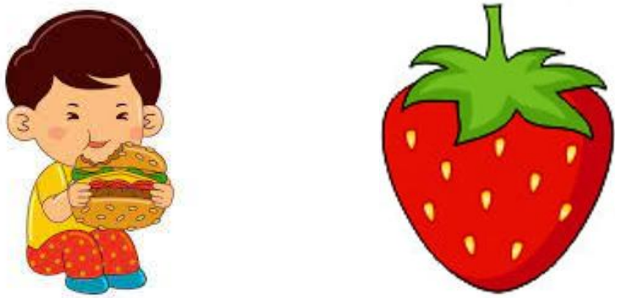
- "Bounding the deviation of a random variable in terms of its variance"

Last Time: Distinct Elements (F_0 Estimation)

- Given a set S of m elements from $[n]$, let f_i be the frequency of element i. (How often it appears)
- Let F_0 be the frequency moment of the vector:

$$F_0 = |\{i : f_i \neq 0\}|$$

- Goal: Given a set S of m elements from $[n]$ and an accuracy parameter ε, output a $(1 + \varepsilon)$-approximation to F_0



Distinct Elements (F_0 Estimation)

- Intuition: How is this done in practice?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers

- Suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$

- How many numbers are in S'?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers
- Suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$
- Can we use S' to get a good estimate of N?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers, suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$

- We have $\mathrm{E}[|S'|] = \frac{N}{2}$ and $\mathrm{Var}[|S'|] \leq \frac{N}{2}$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \ldots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \ldots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$
- Let $X = X_1 + \cdots + X_N$, so that $X = |S'|$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \ldots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$
- Let $X = X_1 + \cdots + X_N$, so that $X = |S'|$
- $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_N] = N \cdot \mathrm{Var}[X_i]$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \dots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$
- Let $X = X_1 + \cdots + X_N$, so that $X = |S'|$
- $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_N] = N \cdot \mathrm{Var}[X_i]$
- $\mathrm{Var}[X_i] = \mathrm{E}\left[X_i^2\right] - \mathrm{E}[X_i]^2$

Distinct Elements (F_0 Estimation)

- Claim: We have $\mathrm{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \ldots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$
- Let $X = X_1 + \cdots + X_N$, so that $X = |S'|$
- $\mathrm{Var}[X] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_N] = N \cdot \mathrm{Var}[X_i]$
- $\mathrm{Var}[X_i] = \mathrm{E}\left[X_i^2\right] - \mathrm{E}[X_i]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

Distinct Elements (F_0 Estimation)

- Claim: We have $\text{Var}[|S'|] \leq \frac{N}{2}$
- Let $X_1, \ldots, X_N$ be indicator random variables so that $X_i = 1$ if the i-th element of S is sampled into S' and otherwise $X_i = 0$
- Let $X = X_1 + \cdots + X_N$, so that $X = |S'|$
- $\text{Var}[X] = \text{Var}[X_1] + \cdots + \text{Var}[X_N] = N \cdot \text{Var}[X_i]$
- $\text{Var}[X_i] = \text{E}\left[X_i^2\right] - \text{E}[X_i]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$
- $\text{Var}[|S'|] = \frac{N}{4}$

Distinct Elements (F_0 Estimation)

- What can we say about $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq t\right]$?
- By Chebyshev's inequality, we have $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq 100\sqrt{N}\right] \leq \frac{1}{10}$

Distinct Elements (F_0 Estimation)

- What can we say about $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq t\right]$?

- By Chebyshev's inequality, we have $\Pr\left[\left||S'| - \frac{N}{2}\right| \geq 100\sqrt{N}\right] \leq \frac{1}{10}$

- With probability at least $\frac{9}{10}$,

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100\sqrt{N}$$

Distinct Elements (F_0 Estimation)

- With probability at least $\frac{9}{10}$,

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100\sqrt{N}$$

- Thus, with probability at least $\frac{9}{10}$,

$$N - 200\sqrt{N} \leq 2|S'| \leq N + 200\sqrt{N}$$

Distinct Elements (F_0 Estimation)

- With probability at least $\frac{9}{10}$,

$$N - 200\sqrt{N} \leq 2|S'| \leq N + 200\sqrt{N}$$

- If $200\sqrt{N} \leq \frac{N}{100}$, then $N - 200\sqrt{N} \leq 2|S'| \leq N + 200\sqrt{N}$ implies

$$0.99N \leq 2|S'| \leq 1.01N$$

- Very good approximation to N

Distinct Elements (F_0 Estimation)

- What algorithm does this suggest?

Distinct Elements (F_0 Estimation)

- What algorithm does this suggest?

- Sample each item of the *universe* with probability $\frac{1}{2}$, acquire new universe U'
- Let S' be the items in the data stream that are in U'
- Output $2|S'|$

Distinct Elements (F_0 Estimation)

- Sample each item of the *universe* with probability $\frac{1}{2}$, acquire new universe U'
- Let S' be the items in the data stream that are in U'
- Output $2|S'|$

- What’s the problem with this approach?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers

- Suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$

- Can we use S' to get a good estimate of N?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers

- Suppose we form set S' by sampling each item of S with probability $\bigcirc\!\!\!\!\!\!p$

- Can we use S' to get a good estimate of N?

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers, suppose we form set S' by sampling each item of S with probability $\frac{1}{2}$
- We have $\mathrm{E}[|S'|] = \frac{N}{2}$ and $\mathrm{Var}[|S'|] \leq \frac{N}{2}$

Distinct Elements (F_0 Estimation)

- Let S be a set of N numbers, suppose we form set S' by sampling each item of S with probability p

- We have $\mathrm{E}[|S'|] = pN$ and $\mathrm{Var}[|S'|] \leq pN$

Distinct Elements (F_0 Estimation)

- (S' is formed by sampling each item of S with probability $\frac{1}{2}$) With probability at least $\frac{9}{10}$,

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100\sqrt{N}$$

- Thus with probability at least $\frac{9}{10}$,

$$N - 200\sqrt{N} \leq 2|S'| \leq N + 200\sqrt{N}$$

Distinct Elements (F_0 Estimation)

- (S' is formed by sampling each item of S with probability p) With probability at least $\frac{9}{10}$,

$$pN - 100\sqrt{pN} \leq |S'| \leq pN + 100\sqrt{pN}$$

- Thus with probability at least $\frac{9}{10}$,

$$N - \frac{100}{\sqrt{p}}\sqrt{N} \leq \frac{1}{p}|S'| \leq N + \frac{100}{\sqrt{p}}\sqrt{N}$$

Distinct Elements (F_0 Estimation)

- (S' is formed by sampling each item of S with probability p) With probability at least $\frac{9}{10}$,

$$N - \frac{100}{\sqrt{p}}\sqrt{N} \leq \frac{1}{p}|S'| \leq N + \frac{100}{\sqrt{p}}\sqrt{N}$$

- If $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$, then $N - \frac{100}{\sqrt{p}}\sqrt{N} \leq \frac{1}{p}|S'| \leq N + \frac{100}{\sqrt{p}}\sqrt{N}$ implies

$$(1-\varepsilon)N \leq \frac{1}{p}|S'| \leq (1+\varepsilon)N$$

Distinct Elements (F_0 Estimation)

- In other words, with probability at least $\frac{9}{10}$, we have that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- What is p?

Distinct Elements (F_0 Estimation)

- In other words, with probability at least $\frac{9}{10}$, we have that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- What is p?
- Recall, we required $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$

Distinct Elements (F_0 Estimation)

- In other words, with probability at least $\frac{9}{10}$, we have that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- What is p?
- Recall, we required $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$, so $p \geq \frac{1000}{\varepsilon^2 N}$

Distinct Elements (F_0 Estimation)

- In other words, with probability at least $\frac{9}{10}$, we have that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- What is p?
- Recall, we required $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$, so $p \geq \frac{1000}{\varepsilon^2 N}$
- What is the problem here?

Distinct Elements (F_0 Estimation)

- In other words, with probability at least $\frac{9}{10}$, we have that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- What is p?
- Recall, we required $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$, so $p \geq \frac{1000}{\varepsilon^2 N}$
- What is the problem here?

Must know N to set p,
but the goal is to find N!

Distinct Elements (F_0 Estimation)

- Observation: We do not need $p = \frac{1000}{\varepsilon^2 N}$, it is also fine to have $p = \frac{2000}{\varepsilon^2 N}$

- How do we find a “good” p?

Finding p

- Observation: We do not need $p = \frac{1000}{\varepsilon^2 N}$, it is also fine to have $p = \frac{2000}{\varepsilon^2 N}$

- How do we find a “good” p?

- What is a “good” p?

Finding p

- What is a “good” p?

- Not too many samples, i.e., S' is small, but enough to find a good approximation to N

- For $p = \Theta\left(\frac{1}{\varepsilon^2 N}\right)$:
  - p is large enough to find a good approximation to N
  - We have $\mathrm{E}[|S'|] = pN = \Theta\left(\frac{1}{\varepsilon^2}\right)$

Finding p

- We want p such that $\mathrm{E}[|S'|] = pN = \Theta\left(\frac{1}{\varepsilon^2}\right)$
- Intuition: Try $p = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots,$ and see which one has

$$\frac{1000}{\varepsilon^2} \le |S'| \le \frac{2000}{\varepsilon^2}$$

- With high probability, one of these guesses will have $\frac{1000}{\varepsilon^2} \le |S'| \le \frac{2000}{\varepsilon^2}$

Finding p

- Intuition: Try $p = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots,$ and see which one has

$$\frac{1000}{\varepsilon^2} \leq |S'| \leq \frac{2000}{\varepsilon^2}$$

- However, the wrong guesses will have too many samples

Finding p

- Intuition: Try $p = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots,$ and see which one has

$$\frac{1000}{\varepsilon^2} \leq |S'| \leq \frac{2000}{\varepsilon^2}$$

- However, the wrong guesses will have too many samples

- Fix: Dynamically changing guess for p and subsampling

Finding p

- Algorithm: Set $U_0 = [n]$ and for each i, sample each element of U_{i-1} into U_i with probability $\frac{1}{2}$
- Start index $i = 0$ and track the number $|S \cap U_i|$ of elements of S in U_i
- If $|S \cap U_i| > \frac{2000}{\varepsilon^2} \log n$, then increment $i = i + 1$
- At the end of the stream, output $2^i \cdot |S \cap U_i|$

Finding p

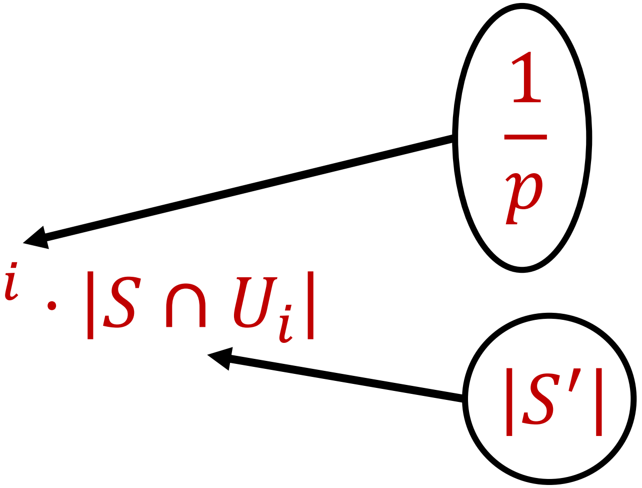
- Algorithm: Set $U_0 = [n]$ and for each i, sample each element of U_{i-1} into U_i with probability $\frac{1}{2}$
- Start index $i = 0$ and track the number $|S \cap U_i|$ of elements of S in U_i
- If $|S \cap U_i| > \frac{2000}{\varepsilon^2} \log n$, then increment $i = i + 1$
- At the end of the stream, output $\underbrace{2^i}_{\boxed{\frac{1}{p}}} \cdot \underbrace{|S \cap U_i|}_{\boxed{|S'|}}$

Finding p

- Recall that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$-approximation of N

- $2^i \cdot |S \cap U_i|$ is a $(1+\varepsilon)$-approximation of N

- At the end of the stream, output $2^i \cdot |S \cap U_i|$



Distinct Elements (F_0 Estimation)

- Summary: Algorithm stores at most $\frac{2000}{\varepsilon^2}\log n$ elements from the stream, uses $\Theta\left(\frac{1}{\varepsilon^2}\log n\right)$ words of space