Define X_i : random variable whose value is the value of the i'th sample

(each X_i is a random element of the array).

Also note: X_i are independent for different i

Expected values:

For each i, $\mathbb{E}[X_i] = \mu$, avg. of elements of the array.

Algorithm's estimate is

$$\hat{X} = \frac{1}{k}\left(X_1 + X_2 + \cdots + X_k\right)$$

By linearity of expectation,

$$\mathbb{E}[\hat{X}] = \mu.$$

Key new computation:

Variance, defined as $V = \mathbb{E}\left[(\hat{x}-\mu)^2\right]$

$$V = \mathbb{E}\left[\left(\frac{X_1+X_2+\dots+X_k}{k}-\mu\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{(X_1-\mu)+(X_2-\mu)+\dots+(X_k-\mu)}{k}\right)^2\right]$$

Expanding gives two kinds of terms:

A) $\mathbb{E}\left[\frac{(X_i-\mu)^2}{k^2}\right]$

B) $\mathbb{E}\left[\frac{(X_i-\mu)(X_j-\mu)}{k^2}\right]$

Due to sampling with replacement, terms of type B are $= 0$!

$$\therefore\ V = \frac{1}{k^2}\sum_{i=1}^{k} \underbrace{\mathbb{E}\left[(X_i-\mu)^2\right]}_{\text{always} \leq 4} \leq \frac{4}{k^2}\cdot k = \frac{4}{k}.$$

<u>Moral:</u>

Variance <u>reduces</u> as # samples (k) <u>increases</u>

Now, what if we use Markov's ineq on

$$V = \mathbb{E}\left[(\hat{X} - \mu)^2\right] ?$$

We get $\Pr\left[(\hat{X} - \mu)^2 > 4 \cdot \frac{4}{k}\right] \leq \frac{1}{4}$

Or equivalently, $\Pr\left[|\hat{X} - \mu| > \frac{4}{\sqrt{k}}\right] \leq \frac{1}{4}$.

This can be generalized (called Chebychev's ineq.)

$$\Pr\left[|\hat{X} - \mu| > \frac{2t}{\sqrt{k}}\right] \leq \frac{1}{t^2}.$$

$\therefore$ w.p. $\geq \frac{3}{4}$, we have $|\hat{X} - \mu| \leq \frac{4}{\sqrt{k}}$

I.e., estimate has error $\leq \frac{4}{\sqrt{k}}$, w.p. $\geq \frac{3}{4}$.

set $k = 16 \cdot 10^4 \rightsquigarrow$ error ≤ 0.01

Moral: more samples $\Rightarrow$ better accuracy!

Variance is a fundamental quantity of random variables.

Prob. value of $3/4$ can be boosted much further.

- Can compute "higher moments"

$$\mathbb{E}\left[(\hat{X}-\mu)^4\right], \ldots$$

- Chernoff bound