CS70
Fall 2013
Lec22

Markov, Chebyshev & Estimating the Bias of a Coin.

$$E[X] = m$$

$$Var[X] = E[(X-m)^2]$$

$$\sigma(X) = \sqrt{Var(X)}$$

<u>Chebyshev</u>: $P[\,|X-m| \geq a\sigma] \leq \frac{1}{a^2}.$

$$P[\,|X-m| \geq \alpha] \leq \frac{Var(X)}{\alpha^2}.$$

Opinion Polling

p = fraction of people

$P[H] = p$.

p unknown.

Estimate $\hat{p}$ for p.

$$\underbrace{|p-\hat{p}| \le \varepsilon}_{\text{accuracy}} \quad \text{with } \sout{\text{confidence}} \text{ probability } \underbrace{\ge 1-\delta}_{\text{confidence}}$$

confidence parameter

Question: For a given ε, δ how many times must we flip coin n.

$S_n = X_1 + X_2 + \cdots + X_n$ = # heads.

$\hat{p} = A_n = \frac{S_n}{n}$.

$E[A_n] = \frac{E(S_n)}{n} = \frac{\cancel{n}E(X_1)}{\cancel{n}} = p$

$$\mathrm{Var}(A_n) = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{\mathrm{Var}(X_1)+\cdots+\mathrm{Var}(X_n)}{n^2}$$

$$= \frac{\cancel{n}\,\mathrm{Var}(X_1)}{n^{\cancel{2}}} = \frac{E(X_1^2)-(E[X_1])^2}{n} = \frac{p-p^2}{n}$$

$$= \frac{p(1-p)}{n}$$

$X_1 = \begin{cases} 1 & \text{wp } p \\ 0 & \text{wp } 1-p \end{cases}$ $\quad E[X_1] = p$

$E(X_1^2) = p$

$\hat{p} = A_n = \dfrac{S_n}{n}$.

$E[A_n] = p$

$Var(A_n) = \dfrac{p(1-p)}{n}$.

$$P[\,\underset{\substack{\| \\ \hat{p}}}{|A_n} - p| \geq \varepsilon] \leq \frac{Var(A_n)}{\varepsilon^2}$$

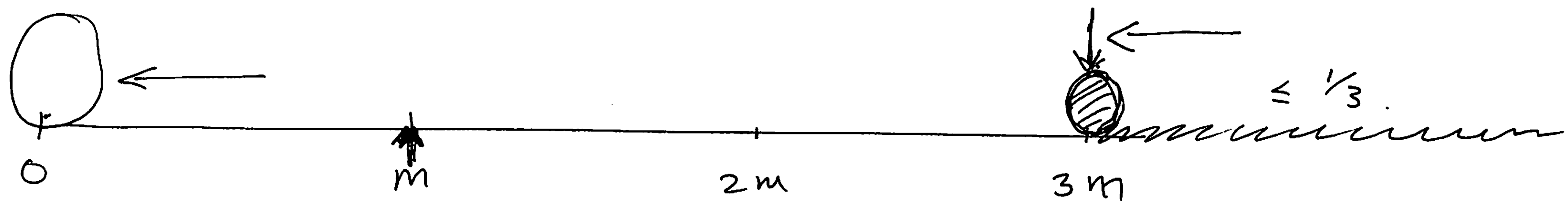
$$= \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2} \leq \delta$$

$$\frac{1}{4n\varepsilon^2} \leq \delta \qquad n \geq \frac{1}{4\varepsilon^2\delta}$$

$\varepsilon = .1 \qquad \delta = .05 \qquad n = 500.$

Markov : $X \geq 0$ $E(X) = m$.
r.v

$$P[X \geq a] \leq \frac{m}{a}.$$



$$E[X] = \sum_{b=0}^{\infty} b\, P[X=b]$$

$$\geq \sum_{b \geq a} b\, P[X=b].$$

$$\geq \sum_{b \geq a} a\, P[X=b]$$

$$= a \sum_{b \geq a} P[X=b]$$

$$= a\, P[X \geq a].$$

Chebyshev: $P[|X-m| \geq a] \leq \dfrac{Var(X)}{a^2}$

$$\iff P[(X-m)^2 \geq a^2] \leq \frac{Var(x)}{a^2}.$$

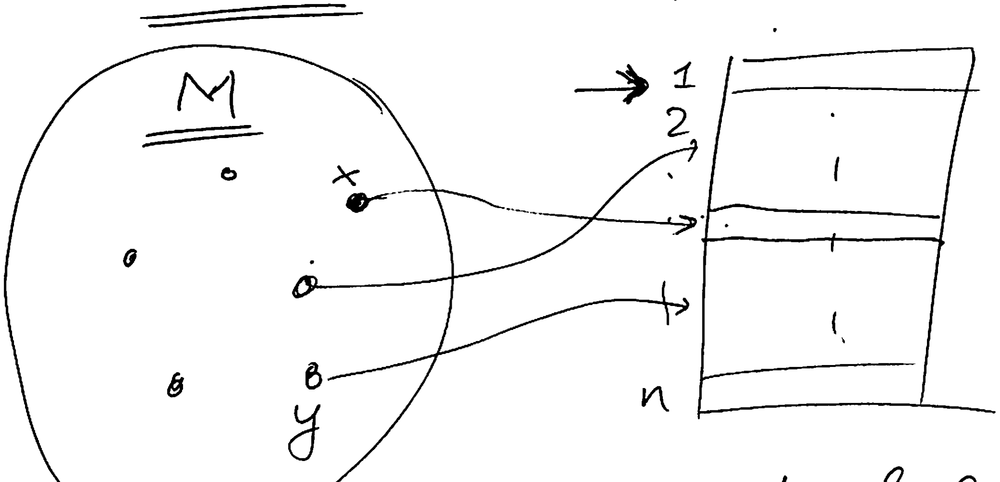
$$Y = (X-m)^2 \qquad Y \geq 0 \qquad E(Y) = Var(X)$$

$$P[Y \geq a^2] \leq \frac{E(Y)}{a^2} = \frac{Var(X)}{a^2}.$$

Hashing

m items $\qquad$ Hash Table $\qquad$ $\boxed{m \approx \cancel{100}n}$ $\boxed{m \simeq n}$

Heavily loaded case.



How many hashed to entry #1?

$$X = X_1 + X_2 + \cdots + X_m. \qquad X_i = \begin{cases} 1 & \text{if } i\text{th item in entry \#1} \\ 0 & \text{o.w.} \end{cases}$$

$$P[X_i = 1] = \frac{1}{n}. \qquad E[X] = m \times \frac{1}{n} = \frac{m}{n}.$$

$$\mathrm{Var}(X) = m\,\mathrm{Var}(X_1) = m \times \frac{1}{n}\left(1 - \frac{1}{n}\right) \le \frac{m}{n}.$$

By Chebyshev $\quad P\left[\left|X - \frac{m}{n}\right| \ge a\right] \le \frac{m}{n\,\boxed{a^2}}$

$m = n$ this is $\frac{1}{a^2}$.

$$\frac{m}{n} = 1$$

$\rightarrow \mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$

true even if X_i are pairwise independent.

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \qquad \text{if } X, Y \text{ indep.}$$

$$E[X \cdot Y] = E[X] \cdot E[Y] \qquad \text{if } X, Y \text{ } \underline{\underline{\text{indep}}}$$

$$\begin{aligned}
E[X \cdot Y] &= \sum_{a,b} a \cdot b \; P[X=a \;\&\; Y=b] \\
&= \sum_{a,b} a \cdot b \; P[X=a] \, P[Y=b] \\
&= \left(\sum_{a} a \, P[X=a]\right)\left(\sum_{b} b \, P[Y=b]\right). \\
&= E(X) \, E(Y).
\end{aligned}$$

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n)$$

$$= E\left[(X_1 + X_2 + \cdots + X_n)^2\right] - \left(E[X_1 + X_2 + \cdots + X_n]\right)^2$$

$$= E\left[\sum_i X_i^2 + 2\sum_{i,j} X_i X_j\right] - \left(\sum_i E[X_i]\right)^2$$

$$\left(E[X_1] + E(X_2) + \cdots + E(X_n)\right)^2$$

$$= \sum_i \underline{E[X_i^2]} + 2\sum_{i,j} E[X_i X_j] - \sum_i \underline{\left(E(X_i)\right)^2} - 2\sum_{i,j} E[X_i]E[X_j]$$

$$= \sum_i \underbrace{\left(E[X_i^2] - \left(E[X_i]\right)^2\right)}_{\mathrm{Var}(X_i)} + 2\sum_{i,j} \underbrace{\cancel{\left(E[X_i X_j] - E[X_i]E[X_j]\right)}}_{\text{pairwise independent}}$$

$$= \sum_i \mathrm{Var}(X_i).$$

Estimating Bias of a Coin

$P[+1] = p$

n times estimate $\hat{p}$.

$|p - \hat{p}| \le \varepsilon$ with prob $\ge 1-\delta$.

Chebyshev: $n \ge \dfrac{1}{4\varepsilon^2\delta}$

Chernoff: $n = \dfrac{O(\log \frac{1}{\delta})}{\varepsilon^2}$ $\qquad \delta \to \log \frac{1}{\delta}$

accuracy is expensive, confidence is cheap.

Law of large Numbers: $X = \underline{X_1} + \cdots + \underline{X_n}$.

mean m and finite variance σ^2

$P\left[\left|\frac{X}{n} - m\right| \ge \varepsilon\right] \longrightarrow 0$ as $n \to \infty$.