11/18/2021

CS70 LECTURE 23 : CONCENTRATION BOUNDS

Consider a nonnegative variable X.
For any a, let Y be a random variable s.t. $Y = \begin{cases} 0 & \text{if } X < a \\ a & X \geq a \end{cases}$

Then $Y \leq X$.
Since $X - Y \geq 0$, $\mathbb{E}[X - Y] \geq 0 \Rightarrow \mathbb{E}[X] - \mathbb{E}[Y] \geq 0 \Rightarrow \mathbb{E}[X] \geq \mathbb{E}[Y]$.

$\mathbb{E}[Y] = 0 \cdot \mathbb{P}[X < a] + a\,\mathbb{P}[X \geq a] \leq \mathbb{E}[X]$
$\Rightarrow a\,\mathbb{P}[X \geq a] \leq \mathbb{E}[X] \Rightarrow \boxed{\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.}$

Markov's Inequality
$$\begin{matrix} X \geq 0 \\ a > 0 \end{matrix} \Rightarrow \mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Chebyshev's Inequality
Special application of Markov's Inequality
Let X be a random variable (not necessarily nonnegative)
Let $\mu = \mathbb{E}[X]$. $\sigma_X^2 \triangleq \mathbb{E}[(X - \mu_X)^2]$.

Chebyshev's inequality states that:
$$\mathbb{P}[\,|X - \mu_X| \geq c\,] \leq \frac{\sigma_X^2}{c^2}$$

Upperbound on how far X can be away from its mean.
i.e. if the variance is small, then the probability that X can be very much away from the mean is smaller.

Proof:
Let $Y = |X - \mu|^2$. Then $Y \geq 0$.
Let $a = c^2$.
Applying Markov's inequality, $\mathbb{P}[Y \geq c^2] \leq \frac{\mathbb{E}[Y]}{c^2}$
$\Rightarrow \mathbb{P}[Y \geq c^2] = \mathbb{P}[\,|X - \mu|^2 \geq c^2] = \mathbb{P}[\,|X - \mu| \geq c] \leq \frac{\mathbb{E}[Y]}{c^2}$
$\therefore \mathbb{P}[\,|X - \mu| \geq c] \leq \frac{\mathbb{E}[Y]}{c^2} = \frac{\mathbb{E}[\,|X - \mu|^2]}{c^2} = \frac{\mathrm{Var}[X]}{c^2}.$

Loose Interpretation for Markov's Inequality
If $\mathbb{E}[X]$ is small, then the probability that X takes on large values of a is small

Sample Mean

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \quad \text{(Sample mean)}$$

μ: $X_i = 1$

$1-\mu$: $X_i = 0$

μ is unknown, but fixed
want to estimate μ.

X_i's are independent, identically distributed (based on the same underlying Bernoulli)
(e.g. n independent flips of a coin)

The sample mean (M_n) is also a random variable in its own right.

$$\mathbb{E}[M_n] = \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n}\underbrace{\sum_{i=1}^{n} \mathbb{E}[X_i]}_{\text{all equal}} = \mathbb{E}[X]$$

all equal since based off same underlying experiment

M_n is an "estimator" for $\mathbb{E}[X]$

$\therefore \boxed{\mathbb{E}[M_n] = \mathbb{E}[X]}$

M_n is an "unbiased estimator" for $\mathbb{E}[X]$

Unbiased estimator: the mean of the estimator is the same as the mean of the unknown quantity.

$$\mathrm{Var}[M_n] = \mathrm{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \left(\frac{1}{n}\right)^2 \mathrm{Var}[X_1 + X_2 + \cdots + X_n]$$

$$= \left(\frac{1}{n}\right)^2 \left(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]\right)$$

$$= \left(\frac{1}{n}\right)^2 \cdot n\,\mathrm{Var}[X] = \frac{\mathrm{Var}[X]}{n} \qquad \therefore \boxed{\mathrm{Var}[M_n] = \frac{\mathrm{Var}[X]}{n}}$$

$$\lim_{n\to\infty} \mathrm{Var}[M_n] = 0$$

As n increases, obtain a better and better estimate of the true mean.

(weak) Law of Large Numbers

$$\mathbb{P}(|M_n - \mathbb{E}[M_n]| \geq \varepsilon) \leq \frac{\mathrm{Var}[M_n]}{\varepsilon^2} \quad \text{(by Chebyshev)}$$

$$\Rightarrow \boxed{\mathbb{P}[|M_n - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}[X]}{n\varepsilon^2}}$$

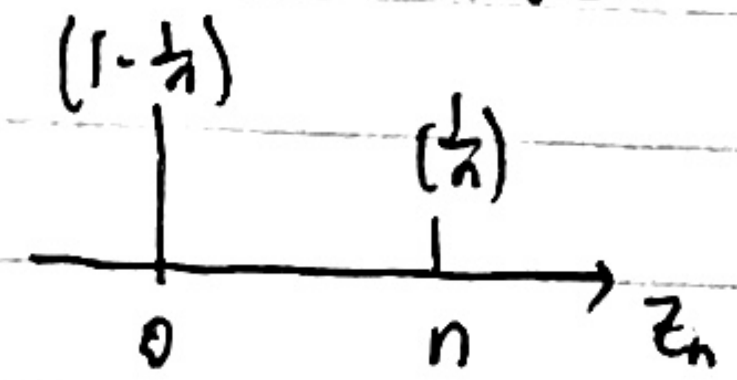
weak means a weak form of convergence

$$\lim_{n\to\infty} \mathbb{P}[|M_n - \mathbb{E}[X]| \geq \varepsilon] \leq \lim_{n\to\infty} \frac{\sigma_X^2}{n\varepsilon^2} = 0$$

Since probability is non negative to begin with, $\lim_{n\to\infty} \mathbb{P}[|M_n - \mathbb{E}[X]| \geq \varepsilon] = 0$ for any arbitrary $\varepsilon > 0$.

|| Sample mean converges in probability to the true mean $\mathbb{E}[X]$ as the number of trials (n) tends to ∞.

Consider a random variable Z_n, with the following probability distribution.



$$\mathbb{E}[Z_n] = 1$$

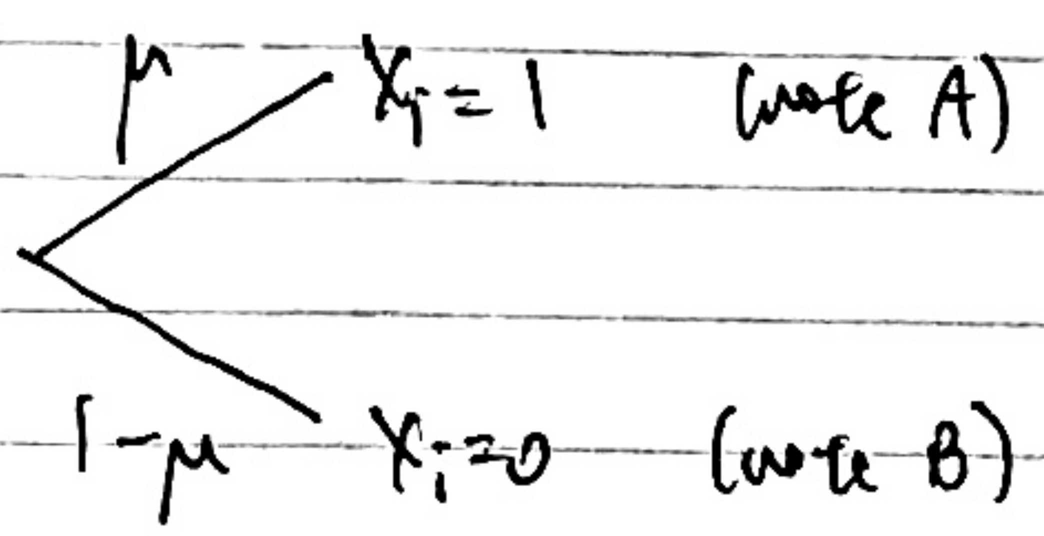
$$Var[Z_n] = \frac{1}{n}(n-1)^2 = \frac{(n-1)^2}{n}.$$

$$\text{As } n \to \infty,\ Var[Z_n] = \lim_{n\to\infty} \frac{(n-1)^2}{n} = \infty$$

Pollster Problem

Want to estimate the true fraction μ $(0 < \mu < 1)$ of US voters who believe in some issue.

Determine the minimum number of voters n that we must poll so that we are at least 95% confident that our estimate M_n is within the range $(\mu - \varepsilon, \mu + \varepsilon)$



$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$ our estimate $\mu = \mathbb{E}[X]$.

Want $\mathbb{P}[\underbrace{|M_n - \mu| < \varepsilon}_{\text{accuracy}}] \geq \underbrace{0.95}_{\text{confidence}}$

Not in the form of Chebyshev's inequality. Just trivially massage.

$$\mathbb{P}[|M_n - \mu| < \varepsilon] = 1 - \mathbb{P}[|M_n - \mu| \geq \varepsilon] \geq 0.95$$

$$\Leftrightarrow \mathbb{P}[|M_n - \mu| \geq \varepsilon] \leq 0.05$$

Know $\mathbb{E}[X_i] = \mu$, $Var[X] = \mu(1-\mu)$

$$\Rightarrow Var[M_n] = \frac{\mu(1-\mu)}{n}$$

Chebyshev's inequality says $\mathbb{P}[|M_n - \mu| \geq \varepsilon] \leq \frac{Var[M_n]}{\varepsilon^2} = \frac{Var[X]}{n\varepsilon^2} \leq 0.05$

Set $\varepsilon = 0.01$, we want $M_n \in [\mu - 0.01, \mu + 0.01]$.

worst case: maximize $\sigma_X^2 = \mu(1-\mu) \overset{\text{AM-GM}}{\leq} \frac{1}{4}$.

$$\Rightarrow \frac{\sigma_X^2}{n(0.01)^2} \leq 0.05 \Rightarrow \boxed{n \geq 50000}$$

To alleviate this large number
- compromise on accuracy (ε)