Markov Inequality

$$\Pr(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

X ↓ nonnegative

Chebyshev Inequality

$$\Pr(|Y - \mu| \geq \alpha) \leq \frac{\mathrm{Var}(Y)}{\alpha^2}$$

Example: $X \sim \mathrm{Bin}(n, \frac{1}{2})$

$$\Pr\left(X \geq \frac{3}{4}n\right) \leq \frac{\frac{n}{2}}{\frac{3n}{4}} = \frac{2}{3}$$ Markov

$$\Pr\left(X \geq \frac{3}{4}n\right) \leq \Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{\mathrm{Var}(X)}{\left(\frac{n}{4}\right)^2} = \frac{\frac{n}{4}}{\left(\frac{n}{4}\right)^2} = \frac{4}{n}$$

$\xrightarrow[n \to \infty]{} 0$

Chebyshev

Sampling & Polling

What fraction of people approve of president?

Poll: call up n random people

$X = X_1 + X_2 + \dots + X_n$

Define $\bar{X} = \frac{X}{n}$ as our estimate

Questions:

What should n be?

How confident are we?

How good an estimate?

can never say my polling estimate 100% guaranteed to be within $\pm 2\%$ of truth.

Question: Given Θ, $1-\varepsilon$: how large does n need to be

so

$$\Pr(|\bar{X} - p| \leq \Theta) \geq 1 - \varepsilon$$

Θ: 0.04 — margin of error

$1-\varepsilon$: 0.95 — confidence

Ex: how big does n need to be so 95% confident that $\bar{X}$ is within 4% of truth (p)?

Want $\Pr(|\bar{X}-p| < 0.04) \geq 0.95$

$\equiv \Pr(|\bar{X}-p| \geq 0.04) \leq 0.05$

$\leq \frac{\mathrm{Var}(\bar{X})}{(0.04)^2}$ by Chebychev

$\leq \frac{1}{(0.04)^2 4n}$

$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X}{n}\right)$

$= \frac{1}{n^2}\mathrm{Var}(X)$

$= \frac{np(1-p)}{n^2}$

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

$\leq \frac{1}{4n}$

Suffices $\frac{1}{(0.04)^2 4n} \leq 0.05$ $\Rightarrow \frac{1}{0.05(0.04)^2 4} \leq n$

$n \geq 3125$

Notes:

- # of samples n doesn't depend on size of total population
- $[p-0.04, p+0.04]$ sometimes called confidence interval

100,000 computers

each indep sends packet w/ prob $q = 0.01$ each secs

Router processes its buffer each sec

How many packet buffers so it drops a packet:

never: 100,000

With prob $\leq 10^{-6}$ each hour?

$X_{it} = \begin{cases} 1 & \text{if computer } i \text{ sends a packet in } t^{th} \text{ second} \\ 0 & \text{otherwise} \end{cases}$

$X_t = \sum_{1 \leq i \leq n} X_{it}$ # of packets sent in a second

B: size of buffer needed so that in T secs prob of overflow $\leq 10^{-6}$

for what B is

$\Pr(\exists t \; 1 \leq t \leq T \text{ s.t. } X_t \geq B) \leq \varepsilon$

First, $\Pr(X_t \geq B)$

By Chernoff bound:

$$\Pr(X > \underbrace{(1+\delta)\mu}_{\#\text{ of buffers } B}) \leq e^{-\frac{\delta^2 \mu}{2}}$$

$$B = (1+\delta)\mu = (1+\delta)(1000)$$

$$\Pr(\text{overflow in } T \text{ secs}) = \Pr(\exists t, 1 \leq t \leq T \text{ s.t. } X_t > (1+\delta)\mu)$$

$$\leq T e^{-\frac{\delta^2 \mu}{2}}$$

$$\leq T e^{-\delta^2 \mu/2}$$

We want $T e^{-\delta^2 \mu/2} \leq \varepsilon$

where $\mu = 1000$

T: # secs in year

$\varepsilon = 10^{-6}$

Recipe: solve for δ

use that to determine $B = (1+\delta)\mu$

$$Te^{-\delta^2\mu/2} \le \varepsilon$$

$$e^{\frac{\delta^2\mu}{2}} \ge \frac{T}{\varepsilon} \qquad \frac{\delta^2\mu}{2} \ge \ln\left(\frac{T}{\varepsilon}\right)$$

$$\delta^2 \ge \frac{2}{\mu}\ln\left(\frac{T}{\varepsilon}\right)$$

$$\delta \ge \sqrt{\frac{2}{\mu}\ln\left(\frac{T}{\varepsilon}\right)}$$

$$B = (1+\delta)\mu$$

Example: $\mu = 1000$

$T = 60 \cdot 60$ (min, sec)

$\varepsilon = 10^{-6}$

$$\delta = \sqrt{\frac{2}{1000}\ln\left(\frac{3600}{10^{-6}}\right)} = 0.2097$$

Buffer size $= 1.2097 \times 1000 \approx 1210$

Twitter processes $\approx$ 600,000,000 tweets per day 6×10^8

Assume tweets independent

chars / tweet $\sim U[10, 140]$

What is $\approx$ prob that twitter processes between 3.899×10^{10} & 3.90001×10^{10} characters? Use CLT

$X = X_1 + X_2 + \ldots + X_n \qquad n = 6 \times 10^8 \qquad X_i \sim U[10, 140]$

$E(X) = n \cdot 80 = 6 \times 10^8 \cdot 75 = 3.9 \times 10^{10}$

$Var(X) = Var(X_1 + X_2 + \ldots + X_n) = n\,Var(X_i) = 6 \times 10^8 \cdot 1430$

$\sigma = 926283$

$$\frac{X - 3.9 \times 10^{10}}{926283} \sim N(0,1)$$