LECTURE 6

Previous class: Hoeffding's inequality:

> If $X_1, \dots, X_N$ are independent r.v's such that $X_i \in [a_i, b_i]$ $\forall i$,
> then $S_N = \sum_{i=1}^{N} X_i$ satisfies
> $$\mathbb{P}\left\{ |S_N - \mathbb{E}S_N| \geq t \right\} \leq 2\exp\left(- \frac{2t^2}{\sum_{i=1}^{N} (b_i - a_i)^2} \right) \quad \forall t \geq 0$$

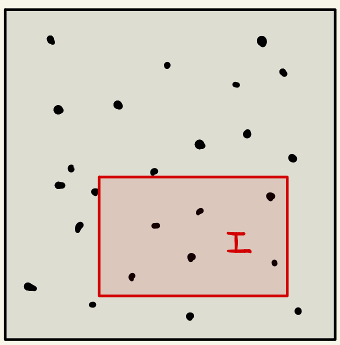
• Example: $X_i \sim \mathrm{Ber}(p) \Rightarrow a_i = 0,\ b_i = 1$; $\mathbb{E}S_N = pN$, $t := \delta N$

Hence $S_N \sim \mathrm{Binom}(N, p)$ satisfies

$$\boxed{\mathbb{P}\left\{ |S_N - pN| \geq \delta N \right\} \leq 2\exp\left(-2\delta^2 N\right)}$$

exponentially small in N. ☺

Today: an application for: DISCREPANCY



• Throw N random points into the square $[0,1]^2$ independently and uniformly.

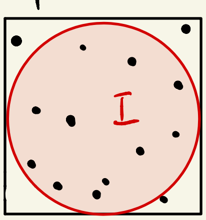
• $\forall$ subset $I \subset [0,1]^2$, expected fraction of pts in I = area (I).

• Why? $N_I := \#(\text{pts in } I) \sim \mathrm{Binom}(N, p_I)$

$$p_I = \mathbb{P}\{\text{a random uniform pt} \in I\} = \text{area}(I)$$

$$\Rightarrow \mathbb{E}N_I = p_I N \quad \Rightarrow \quad \boxed{\mathbb{E}\left[\frac{N_I}{N}\right] = p_I = \text{area}(I)}$$

• Application: a probabilistic computation of π:



$$\frac{N_I}{N} \xrightarrow{LLN} \mathbb{E}\left[\frac{N_I}{N}\right] = \text{area}(I) = \frac{\pi}{4}$$

• Mean squared error: $\mathbb{E}\left(\frac{N_I}{N} - p_I\right)^2 = \mathrm{Var}\left(\frac{N_I}{N}\right) = \frac{1}{N^2}\mathrm{Var}(\underbrace{N_I}_{\mathrm{Binom}(N, p_I)}) = \frac{Np(1-p)}{N^2} \leq \frac{1}{N}$

$\Rightarrow$ st. dev $\leq 1/\sqrt{N}$. Chebyshev $\Rightarrow$

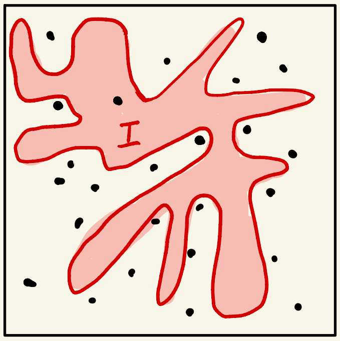
$$\boxed{\mathbb{P}\left\{ \left| \frac{N_I}{N} - \underbrace{p_I}_{\text{area}(I)} \right| = O(1/\sqrt{N}) \right\} \geq 0.99 \quad \forall I \subset [0,1]^2} \qquad (*)$$

• Q: does (*) hold for all I simultaneously, i.e. is it true that

$$P\left\{ \forall I \subset [0,1]^2 : \left| \frac{N_I}{N} - \text{Area}(I) \right| = O(1/\sqrt{N}) \right\} \geq 0.99 \,? \qquad (**)$$

(is there a "universal sample"?)

• No:



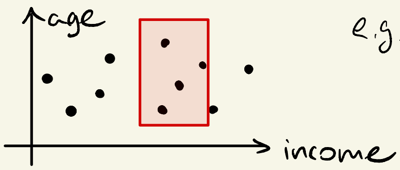
• Q: Does this hold for simple shapes, such as I ∈ rectangles? YES

THM (Discrepancy) A set of N independent random points, uniformly drawn from the square $[0,1]^2$, satisfies the following with probability ≥ 0.99.

For any axis-aligned rectangle $I \subset [0,1]^2$, the fraction of the points in I satisfies

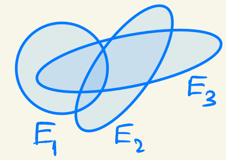
$$\left| \frac{N_I}{N} - \text{area}(I) \right| \leq C \sqrt{\frac{\log N}{N}}$$

• Relevance for statistic: representative sampling:



e.g. We want a sample in which all age brackets and income brackets are fairly represented

PROOF "An epsilon-net method" based on a union bound



$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \le \sum_{i=1}^{\infty} P(E_i) \quad \forall \text{ events } E_i$$

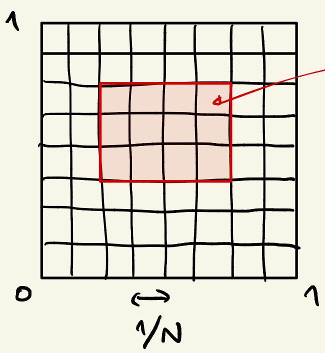
Want to show:

$$P\Big\{ \exists \text{ rectangle } I \;\; \underbrace{\left|\frac{N_I}{N} - p_I\right| > C\sqrt{\frac{\log N}{N}}}_{\overset{!!}{=}\, E_I} \Big\} \le 0.01$$

$$= P\Big(\bigcup_{I \in \text{rectangles}} E_I\Big) \le \sum_{I \in \text{rectangles}} P(E_I)$$

↑ infinite sum ☹ ⇒

① Discretize :



a "grid rectangle"

There are $\le N^4$ grid rectangles. Not ∞ anymore! ☺

② Concentration: $\forall$ fixed grid rectangle I, $N_I \sim \text{Binom}(N, p_I) \Rightarrow$

$$P\left\{ \left|\frac{N_I}{N} - p_I\right| \ge \delta \right\} = P\{ |N_I - p_I N| \ge \delta N \}$$

$$\overset{\text{Hoeffding p. 2}}{\le} 2\exp(-2\delta^2 N) \qquad \forall \delta \ge 0$$

$$\le \frac{1}{100 N^4} \quad \text{if we choose} \quad \boxed{\delta = C\sqrt{\frac{\log N}{N}}}$$

C: a large absolute constant

③ Union Bound:

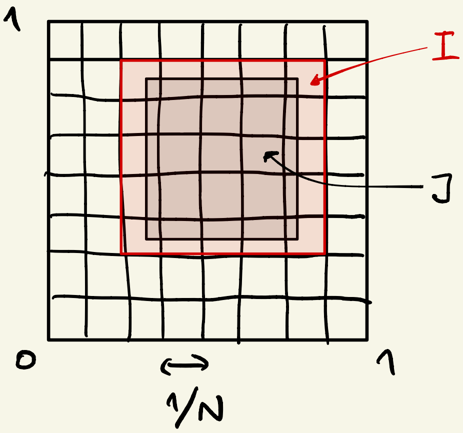
$$P\left\{ \exists I \in \text{grid rectangles}: \left|\frac{N_I}{N} - p_I\right| \ge \delta \right\} \overset{\text{union Bd}}{\le} \sum_{I \in \text{grid rec}} P\left\{ \left|\frac{N_I}{N} - p_I\right| \ge \delta \right\}$$

$$\le N^4 \cdot \frac{1}{100 N^4} = 0.01.$$

We proved: $P\left\{ \forall \text{ grid rec } I: \left|\frac{N_I}{N} - p_I\right| \le C\sqrt{\frac{\log N}{N}} \right\} \ge 0.99.$ ☺

Assume this event occurs

④ Approximation "$\forall$ rectangle $\approx$ a grid rectangle":



- $\forall$ rectangle J lies in a grid rectangle I with area $P_I \le P_J + \frac{4}{N}$. $\Rightarrow$

$$\frac{N_J}{N} \le \frac{N_I}{N} \le P_I + C\sqrt{\frac{\log N}{N}} \quad \text{(by step 3)}$$

$$\le \underbrace{P_J + \frac{4}{N}}_{\text{smaller}} + \underbrace{C\sqrt{\frac{\log N}{N}}}_{\text{larger}} \le P_J + C'\sqrt{\frac{\log N}{N}}.$$

- Similarly, $\frac{N_J}{N} \geq P_J - C'\sqrt{\frac{\log N}{N}}$ (DIY)

Thus:

$$\left|\frac{N_J}{N} - P_J\right| \le C'\sqrt{\frac{\log N}{N}} \quad \forall \text{ rectangle } I. \qquad \text{QED}.$$

REMARKS

① $\log N$ can be removed.

② Uniform distr. on $[0,1]^2$ can be replaced with $\forall$ distr. on $\mathbb{R}^2$

③ Rectangles can be replaced by other simple shapes such as triangles, circles, ellipses...

④ The result can be extended to $\mathbb{R}^d$:

$$\left|\frac{N_I}{N} - P_I\right| \le C\sqrt{\frac{d}{N}} \quad \forall \text{ box } I$$

All these will follow from general VC theory (covered later).