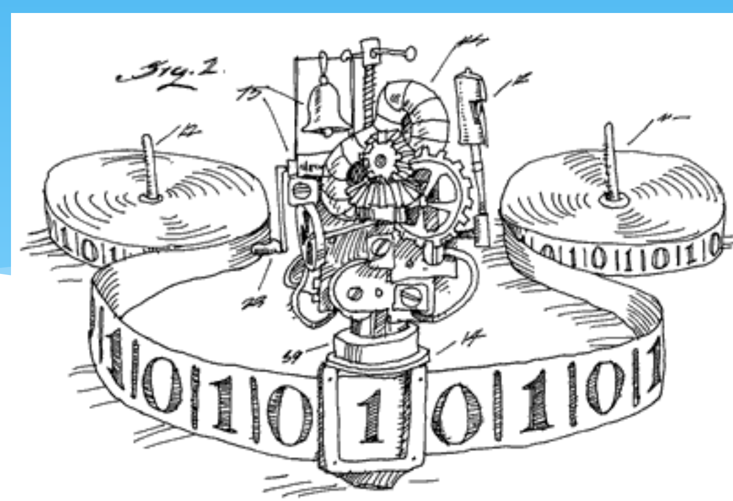
EECS 376: Foundations of Computer Science

Chris Peikert

3 April 2023



Today’s Agenda

* Chernoff-Hoeffding and union bounds
* Polling
* Distinguishing biased coins

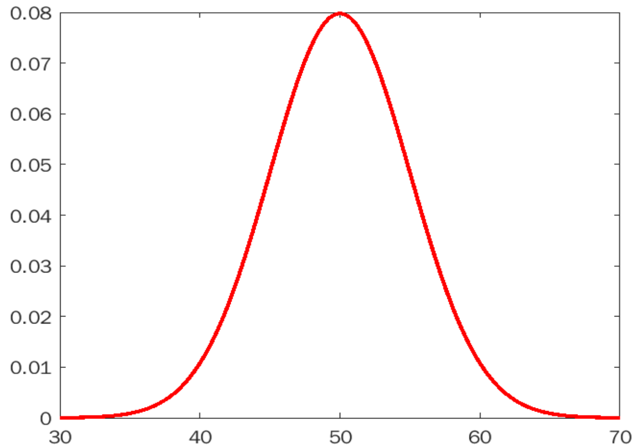
90%+ (??) of randomized algorithms analyses use Chernoff bound + Union bound
EECS 572: Randomness and Computation for many applications

How Many Heads?

- We want to determine if a coin is fair or not.
- **Q:** How "*suspicious*" should we be if we flip it n times and see k heads, for the following values of n and k?
  - $n = 100,\ k = 51$
  - $n = 10{,}000,\ k = 5{,}100$
  - $n = 1{,}000{,}000,\ k = 510{,}000$
- Want to estimate $\Pr[X \geq k]$, where X is the number of heads after flipping a fair coin n times.

Normal Distribution

- A ***normal distribution*** has a "*bell-curve*" shape and is characterized by two parameters: *mean* and *standard deviation*.
  - **Examples:** Height, exam scores, measurement errors are "normal-like"…
- ***66-95-99.7 rule:*** ≈66 , 95 , 99.7% of the area under the curve (i.e., probability) is within 1 , 2 , 3 SD of the mean, respectively.

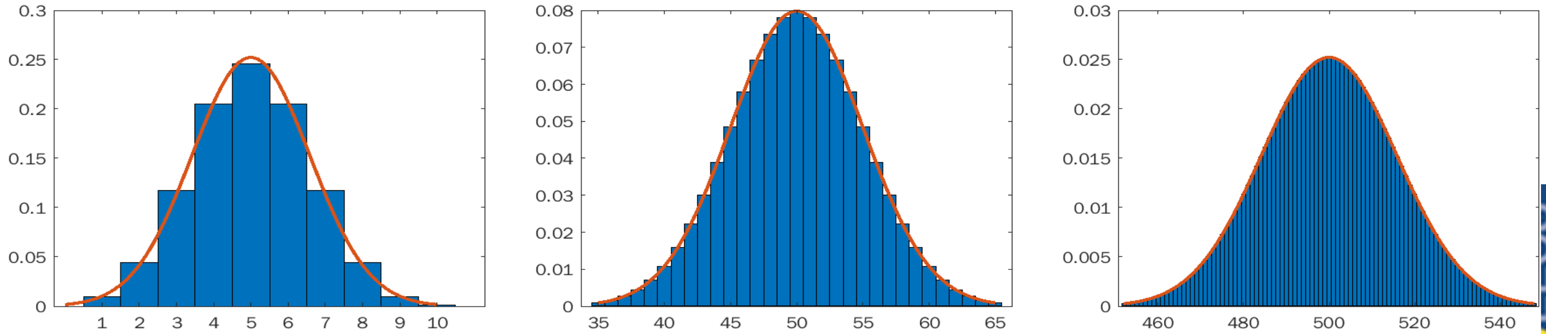


mean 50, SD 5

Note: Distribution is from $-\infty$ to ∞; nothing's 0 here.

Central Limit Theorem

- **(Informal)** For large n, the sum $X = X_1 + \ldots + X_n$ of n ***i.i.d.*** RVs approaches a normal with mean $\mathbb{E}[X] = n\mathbb{E}[X_i]$ and SD=$\sqrt{n}$SD(X_i).
- **Example:** The number of heads after flipping a fair coin n times approaches a normal distribution with mean $n/2$ and SD $\sqrt{n}/2$.
- **Normalize:** use X/n: mean $\mathbb{E}[X/n] = \mathbb{E}[X_i]$ and SD=SD$(X_i)/\sqrt{n}$.



$n = 10$ mean 5, SD 1.58

$n = 100$ mean 50, SD 5

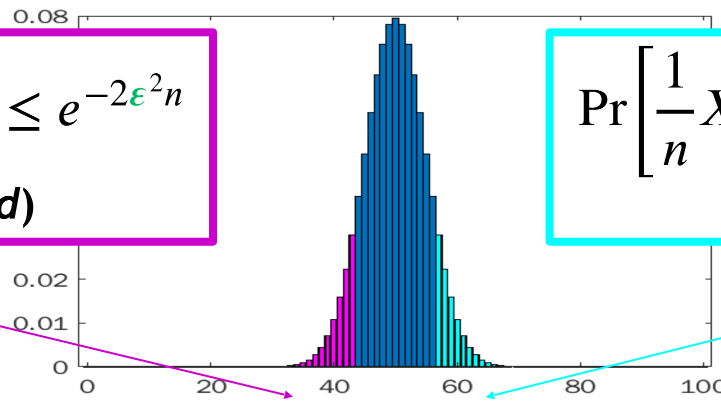
$n = 1000$ mean 500, SD 15.8

Chernoff-Hoeffding Bounds

- If $X = X_1 + X_2 + \ldots + X_n$ is the sum of n ***i.i.d.*** RVs with **each** $\boldsymbol{X_i} \in [\mathbf{0}, \mathbf{1}]$, then, for any $\varepsilon > 0$:

$$\Pr\left[\frac{1}{n}X \le \mathbb{E}[X_i] - \varepsilon\right] \le e^{-2\varepsilon^2 n}$$

(lower tail bound)



$$\Pr\left[\frac{1}{n}X \ge \mathbb{E}[X_i] + \varepsilon\right] \le e^{-2\varepsilon^2 n}$$

(upper tail bound)

Example: flip a fair coin $n = 1{,}000{,}000$ times. Probability we see $\ge (50 + \mathbf{1})\%$ heads is at most $e^{-2(\mathbf{0.01})^2 \cdot 1{,}000{,}000} = e^{-200}$.

- For n flips, we get $\ge k\,\mathbf{SD}(X) = k \cdot \frac{1}{2}\sqrt{n}$ *more* (similarly, *fewer*) heads than mean (= $n/2$) with probability $\le e^{-k^2/2}$.

Polling

- There are m candidates for president. How can we estimate their rates of support ***without*** asking the entire population?
- **A:** Sample people at random and compute the relative frequencies.
- Two types of "accuracy":
  1. How close our estimate could be to the "true" rate
  2. The probability that our estimate is that close
- **Fine print:** "This poll has been conducted with a ***confidence level*** of **95%** and ***margin of error*** $\pm$**2%**"

EECS M ECE CSE

Polling

- **Algorithm for one candidate (approval rating):**
  - Sample a random person, n times ("Do you support Smith?" Yes/No)
  - Let X be the number of supporters
  - Return X/n as the estimate
- Let $0 \leq p \leq 1$ be the true level of support. What should n be to get good "accuracy" with high "confidence"?
- **Fine print:** "This poll has been conducted with a ***confidence level*** of **95%** and ***statistical error*** of $\pm$**2%**"
- Thus, we want $\Pr\left[\left|\frac{1}{n}X - p\right| \leq 0.02\right] \geq 0.95$.

Combined Chernoff-Hoeffding

$$\Pr\left[\frac{1}{n}X \le \mathbb{E}[X_i] - \varepsilon\right] \le e^{-2\varepsilon^2 n}$$

(lower tail bound)

$$\Pr\left[\frac{1}{n}X \ge \mathbb{E}[X_i] + \varepsilon\right] \le e^{-2\varepsilon^2 n}$$

(upper tail bound)

- We want $\Pr\left[\left|\frac{1}{n}X - p\right| \le 0.02\right] \ge 0.95$.
- Define indicators for $i = 1..n$:

$$X_i = \begin{cases} 1, & \text{person } i \text{ supports the candidate} \\ 0, & \text{otherwise} \end{cases}$$

- Then $\mathbb{E}[X_i] = \Pr[X_i = 1] = p$ and $X = X_1 + X_2 + \ldots + X_n$.
- **Q:** What should the value of n be to satisfy the fine print?
- ***Combined CH bound:*** $\Pr\left[\left|\frac{1}{n}X - \mathbb{E}[X_i]\right| \ge \varepsilon\right] \le 2e^{-2\varepsilon^2 n}$

(since $\Pr\left[\left|\frac{1}{n}X - \mathbb{E}[X_i]\right| \ge \varepsilon\right] = \Pr\left[\left(\frac{1}{n}X \le \mathbb{E}[X_i] - \varepsilon\right) \cup \left(\frac{1}{n}X \ge \mathbb{E}[X_i] + \varepsilon\right)\right]$,
and the ***union bound***: $\Pr[A \cup B] \le \Pr[A] + \Pr[B]$)

Polling Analysis

$$\Pr\left[\left|\frac{1}{n}X - \mathbb{E}[X_i]\right| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 n}$$

(combined bound)

* We want $\Pr\left[\left|\frac{1}{n}X - p\right| \leq 0.02\right] \geq 0.95$.
* **Equivalently:** We want $\Pr\left[\left|\frac{1}{n}X - p\right| > 0.02\right] \leq 0.05$.
* By the combined CH bound:

$$\Pr\left[\left|\frac{1}{n}X - p\right| > 0.02\right] \leq \Pr\left[\left|\frac{1}{n}X - p\right| \geq 0.02\right] \leq 2e^{-2 \cdot 0.02^2 n}$$

So, we want $2e^{-2\cdot 0.02^2 n} \leq 0.05$

$$\iff 40 \leq e^{2\cdot 0.02^2 n} \iff \ln(40) \leq 0.0008n \Longleftarrow 4612 \leq n.$$

* **Observe:** n does not depend on the population size!

Polling Multiple Candidates

- **Algorithm for m candidates:**
  - Sample a random person, n times (ask: “Whom do you support?”)
  - Let $X^{(j)}$ be the number of supporters of candidate j
  - For each j, return $X^{(j)}/n$
- **Fine print:** “This poll has been conducted with a ***confidence level*** of $1-\delta$ and ***statistical error*** of $\pm\varepsilon$”
- **Formally:** Let $p_1, \ldots, p_m$ be the support levels of the candidates.
- **We want:** $\Pr\left[\forall j : \left|\frac{X^{(j)}}{n} - p_j\right| \leq \varepsilon\right] \geq 1-\delta.$

Polling Multiple Candidates

- **We want:** $\Pr\left[\forall\, j : \left|\frac{X^{(j)}}{n} - p_j\right| \le \varepsilon\right] \ge 1 - \delta.$
- For $m = 1$, we took $n \ge \ln(2/\delta)\,/\,(2\varepsilon^2)$.
- How many samples should we take now?
- **Wrong answer:** for m candidates we need $n \ge m \cdot \ln(2/\delta)\,/\,(2\varepsilon^2)$.
- **Right answer:** $n \ge \ln(2m/\delta)\,/\,(2\varepsilon^2)$ suffices! (*Log* dep on m.)
- **Proof:** for each $j = 1, \ldots, m$, $\Pr[X^{(j)}$ bad$] \le \delta/m$.

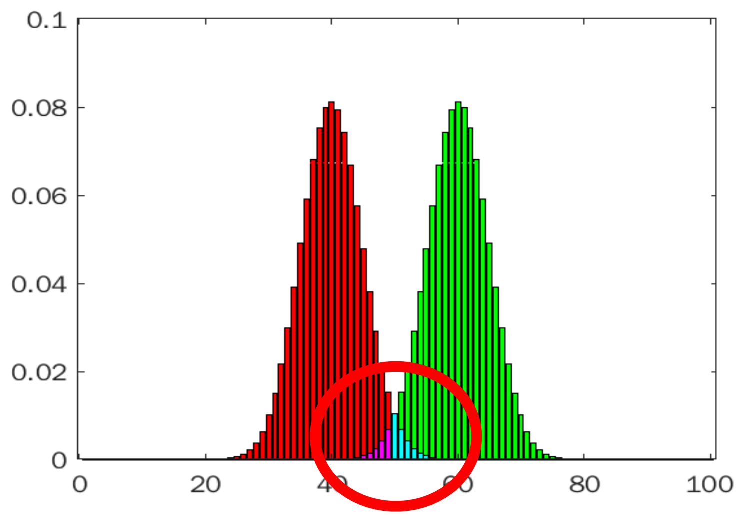
  By union bound, $\Pr[\exists\, j : X^{(j)}$ bad$] \le \sum_j \Pr[X^{(j)}$ bad$] \le \delta$.

Distinguishing Biased Coins

* You're given a coin that is ε-biased to either heads or tails.
  * i.e., $\Pr[H] = \frac{1}{2} + \varepsilon$ **or** $\Pr[H] = \frac{1}{2} - \varepsilon$
* To determine which way it's biased, you flip the coin n times.
  * If you see at least $\frac{1}{2}n$ heads, you guess "H"
  * Otherwise, you guess "T".

Note: We have ***two-sided*** error; ***false positives*** and ***false negatives***!

Q: How large should n be to guarantee an error probability $\leq \delta$?



Probability of False Negatives

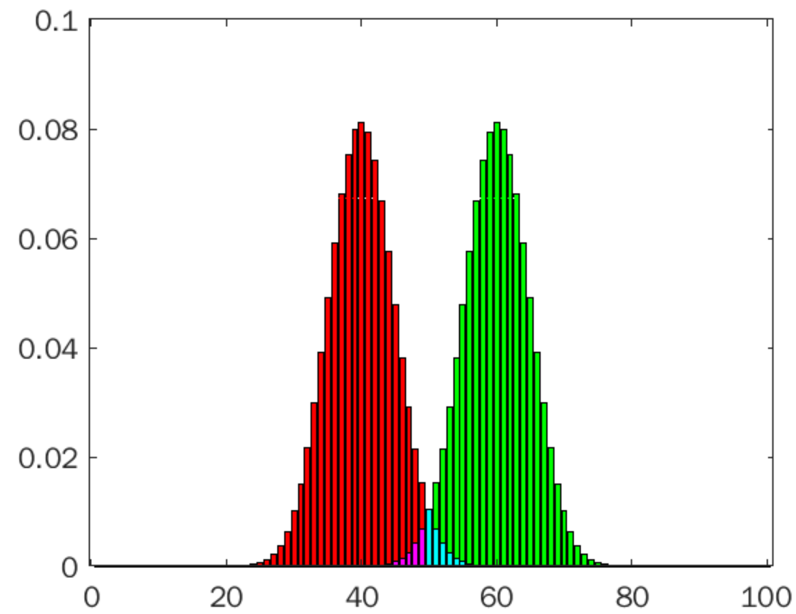
If $X = X_1 + X_2 + \ldots + X_n$ is the sum of n ***i.i.d.*** RVs with **each** $\boldsymbol{X_i} \in [\mathbf{0}, \mathbf{1}]$, then, for any $\varepsilon > 0$:

$$\Pr\left[\frac{1}{n}X \le \mathbb{E}[X_i] - \boldsymbol{\varepsilon}\right] \le e^{-2\varepsilon^2 n}$$

(***lower tail bound***)

- Let X_i be the indicator RV for whether i'th coin flip was H.
- Suppose the coin is ε-biased towards **heads**.
  - Then $\mathbb{E}[X_i] = \frac{1}{2} + \varepsilon$.
- **Q:** When do we get an error (**false negative**) in this case?
- **A:** When $\frac{1}{n}X < \frac{1}{2} = \mathbb{E}[X_i] - \varepsilon$
- Therefore:
  $\Pr[\text{error} \mid H \text{ bias}]$

$$= \Pr\left[\frac{1}{n}X < \mathbb{E}[X_i] - \varepsilon\right] \le e^{-2\varepsilon^2 n}$$



Probability of False Positives

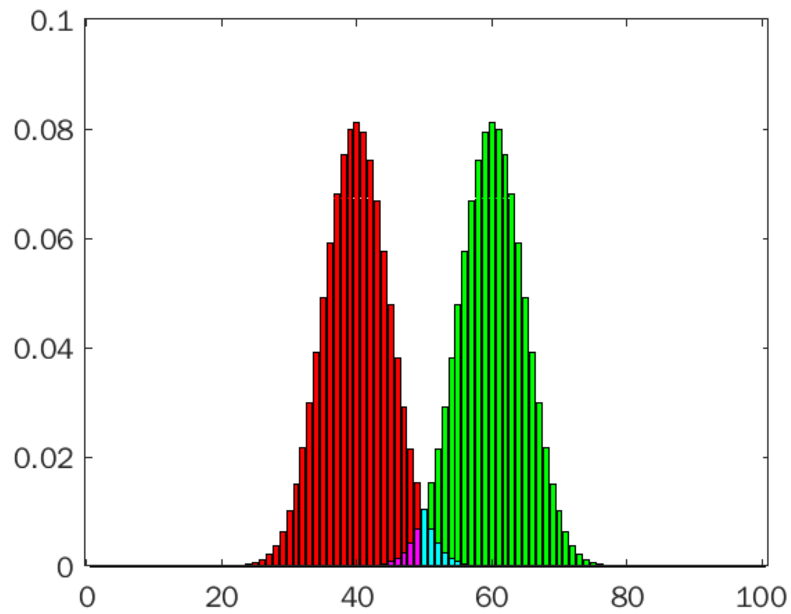
If $X = X_1 + X_2 + \ldots + X_n$ is the sum of n ***i.i.d.*** RVs with **each** $\boldsymbol{X_i} \in [\mathbf{0}, \mathbf{1}]$, then, for any $\varepsilon > 0$:

$$\Pr\left[\frac{1}{n}X \geq \mathbb{E}[X_i] + \boldsymbol{\varepsilon}\right] \leq e^{-2\varepsilon^2 n}$$

(*lower tail bound*)

- Let X_i be the indicator RV for whether i'th coin flip was H.
- Suppose the coin is ε-biased towards **tails**.
  - Then $\mathbb{E}[X_i] = \frac{1}{2} - \varepsilon$.
- **Q:** When do we get an error (**false positive**) in this case?
- **A:** When $\frac{1}{n}X \geq \frac{1}{2} = \mathbb{E}[X_i] + \varepsilon$
- Therefore: $\Pr[\text{error} \mid T \text{ bias}]$

$$= \Pr\left[\frac{1}{n}X \geq \mathbb{E}[X_i] + \varepsilon\right] \leq e^{-2\varepsilon^2 n}$$



What Should n Be?

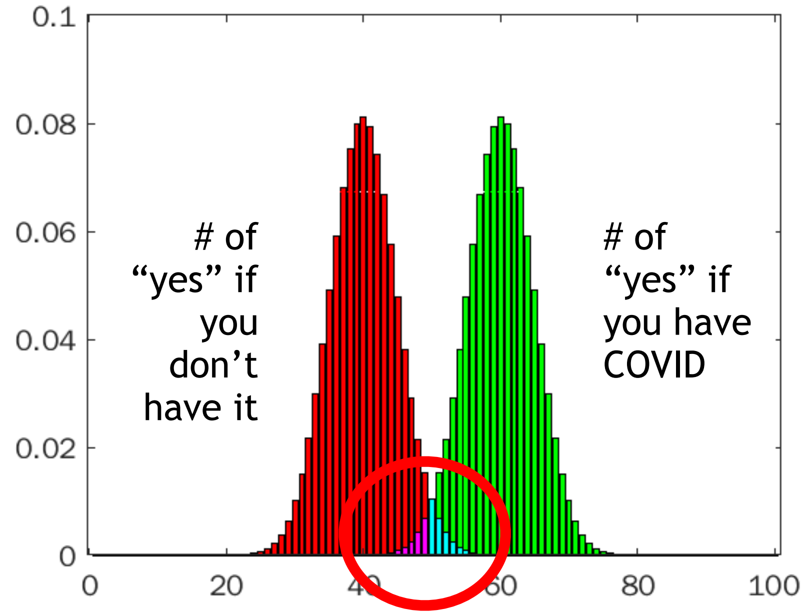
- By previous analysis, in either case, $\Pr[\text{error}] \leq e^{-2\varepsilon^2 n}$.
- What should n be if we want error to be $\leq \delta$?
  - $e^{-2\varepsilon^2 n} \leq \delta \Leftrightarrow n \geq \dfrac{\ln(1/\delta)}{2\varepsilon^2}$.
- **Example:** If $\varepsilon = 0.01$ and $\delta = 0.0001$ (correct 99.99%), then $n \geq \dfrac{\ln\left(0.0001^{-1}\right)}{2 \cdot 0.01^2} \approx 46,052$ flips suffices.
- **Q:** What if we had a fair coin vs 0.01-biased coin (H or T)?
  - Set threshold for guessing fair vs. biased to $0.505n$. Only distance between the means matters for the analysis!
  - Distance halved; n quadruples: $\approx 184,206$

EECS ECE CSE

Extra

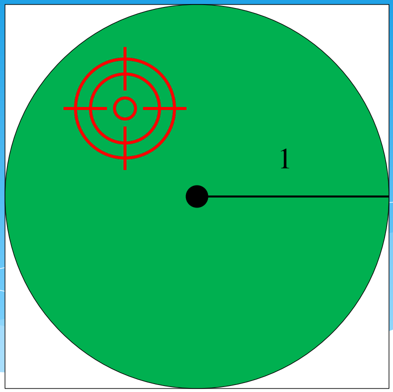
Decreasing Error

- A low-quality COVID test has ***two-sided*** error:
  - If a person has COVID, it says "yes" w.p. 2/3.
  - Otherwise, it says "no" w.p. 2/3.
  - Different runs are independent.
- You decide to buy and run the test n times and take the *majority* answer you get.
- **Q:** How large should n be to guarantee that the answer is correct w.p. $1-\delta$?
  - Same as distinguishing ε-biased coins with $\varepsilon = 1/6$!



Note: *false positives* and *false negatives* are possible!

Estimating π



- Suppose there is a 2x2 square with a unit circle inside
- **Q:** If we toss a dart *uniformly at random* towards the square, what's the probability that we hit the circle?
  - (area of circle)/(area of board) = $\pi/4$
- We toss n darts uniformly at random towards the square
  - X_i = indicator 0/1 RVs for whether we hit circle on i'th toss
  - **Q:** What is $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$?
- **Q:** How might we estimate π by tossing darts?
  - It's *roughly* 4*fraction of times we hit circle; CH to bound error.

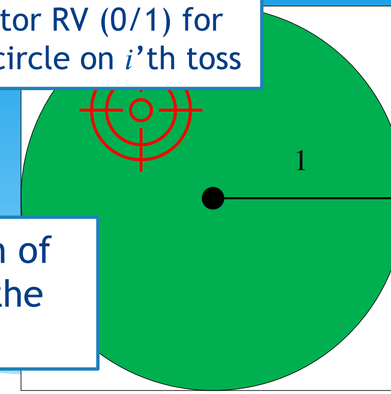
$$\Pr\left[\left|\frac{1}{n}X - \mathbb{E}[X_i]\right| \geq \varepsilon\right] \leq 2e^{-2\varepsilon^2 n}$$

(combined bound)

Math

(to show that this is fairly inefficient)

We toss n darts uniformly at random towards the square
Let X_i = indicator RV (0/1) for whether we hit circle on i'th toss



$\pi \approx$ 4*fraction of times we hit the circle

- Let $X = X_1 + \ldots + X_n$ be the number of times we hit the circle.
- $\mathbb{E}[X_i] = \dfrac{\pi}{4}$ so $\left|\dfrac{1}{n}X - \dfrac{\pi}{4}\right| < \varepsilon$ with probability $\geq 1 - 2e^{-2\varepsilon^2 n}$
- To estimate π within γ, i.e. $\left|\dfrac{4}{n}X - \pi\right| < \gamma$, set $\varepsilon = \dfrac{\gamma}{4}$.
  - $\left|\dfrac{4}{n}X - \pi\right| < \gamma \Leftrightarrow \left|\dfrac{1}{n}X - \dfrac{\pi}{4}\right| < \dfrac{\gamma}{4} = \varepsilon$

    (with probability $\geq 1 - 2e^{-2\varepsilon^2 n} = 1 - 2e^{-\gamma^2 n/8}$)
- For probability $\geq 1 - \delta$, set $n = 8\ln(2/\delta)/\gamma^2$.
  - $1 - 2e$

Example: To get our estimate between 3.140 and 3.142 ($\gamma = 0.001$) 99.99% of the time ($\delta = 0.0001$), we should toss $n \approx 79{,}227{,}901$ darts