Randomized Algorithms 2026 !

This class:
- probability + linear algebra
- data science at scale
- beautiful math

challenging course!

Prereqs: linear algebra, probability, algorithms

Time Expectation: 9.5 hr/week

2.5	1	1	2	1	1	1
class	Problem Set			OH	discussion	"rock"

- Earlier = better
- More drawing = better
- More talking = better

www.rtealwitter.com/rads2026

- discord for communication
- reading for each lecture
- these slides are online

Plan:
- Math Review
- Streaming and Sketching
- Linear Algebra and Spectral Methods

Course Goal: Use randomness to solve problems more efficiently

Probability

Random variable X

e.g., X = outcome of fair dice

For $x \in \{1,2,3,4,5,6\}$,

$\Pr(X=x) = 1/6$

What is X on average?

Expectation of X:

$$\mathbb{E}[X] = \sum_x x \Pr(X=x)$$

$$= \sum_{x=1}^{6} x \, 1/6$$

$$= 1/6\,(1+2+\ldots+6)$$

$$= 21/6$$

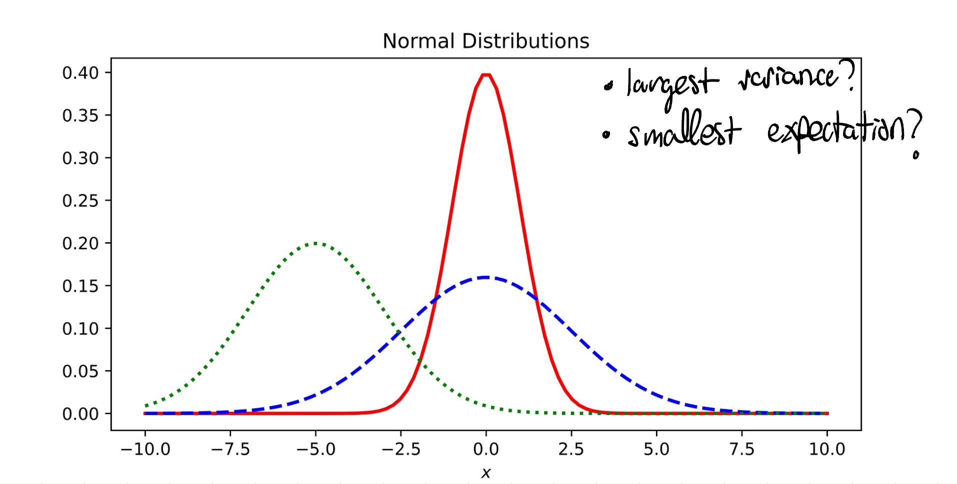
How close is X to its expectation?

Variance of X:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$$

$$= \sum_x (x-\mathbb{E}[X])^2 \Pr(X=x)$$

$$=$$



Multiplying by Scalars

$$\mathbb{E}[\alpha X] = \sum_x \alpha x \Pr(X=x)$$

$$= \alpha \sum_x x \Pr(X=x)$$

$$= \alpha \mathbb{E}[X]$$

$$\mathrm{Var}(\alpha X) =$$

Independent Random Variables

A, B are events

$A =$ dice shows 1 or 2

$B =$ dice roll is odd

"and" "given"

$$Pr(A \cap B) = Pr(A)\, Pr(B \mid A)$$
$$= Pr(B)\, Pr(A \mid B)$$

"if and only if"

$Pr(B \mid A) = Pr(B)$ iff A, B independent

$\Leftrightarrow$ $Pr(A \cap B) = Pr(A)\, Pr(B)$ iff A, B independent

Example A, B indep?

Random Variables X, Y are indep

iff $Pr(X = x \cap Y = y) = Pr(X = x)\, Pr(Y = y)$

for all possible outcomes x, y

Independent?

- Heads on 1st flip, Heads on 2nd
- Ace on 1st draw, Ace on 2nd
- Rain, carrying an umbrella

Linearity of Expectation

Thm: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Proof: $\mathbb{E}[X+Y] = \sum_x \sum_y (x+y) \Pr(X=x \cap Y=y)$

$$= \sum_x \sum_y x \Pr(X=x \cap Y=y) + \sum_x \sum_y y \Pr(X=x \cap Y=y)$$

$$= \sum_x x \sum_y \Pr(X=x \cap Y=y) + \sum_y y \sum_x \Pr(Y=y \cap X=x)$$

$$= \sum_x x \Pr(X=x) + \sum_y y \Pr(Y=y)$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Expected number of shared birthdays?

Week 1 Thursday

- No OH Monday 1/26
  - ⇒ Problem 2 due 2/2
- Join discord!
- Discussions!
  - ↳ host can write quiz question

Today:

- More expectation/variance
- Set size estimation!

Expectation & Variance

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

- never?
- sometimes?
- always?

If X, Y indep, $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

$$\mathbb{E}[XY] = \sum_{x,y} xy \Pr(X=x \cap Y=y)$$

indep

$$= \sum_x \sum_y x\, y \Pr(X=x)\Pr(Y=y)$$

$$= \sum_x x \Pr(X=x) \sum_y y \Pr(Y=y)$$

$$= \mathbb{E}[X]\,\mathbb{E}[Y]$$

Dependent Example

$X = \begin{cases} 1 & \text{if tails} \\ 0 & \text{if heads} \end{cases}$ $\quad Y = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$

$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

If X, Y indep, $Var(X+Y) = Var(X) + Var(Y)$

New person writes please!

Set Size Estimation

Problem:

Given *uniform* query access to a set, estimate the size of the set.

- Ecology
- social networks
- Internet indexing

Suppose Wikipedia claims it has $n = 1{,}000{,}000$ articles, how do we check?

Naive Solution: Keep querying until we get n unique elements

The issue is that the expected queries to find all n is $n \log n$

Clever Solution: Count duplicates!

Intuitively, more duplicates $\rightarrow$ smaller n

Coupon Collector Problem

$$\frac{\partial}{\partial x} \sum_{k=0}^{\infty} x^k = \frac{\partial}{\partial x} \frac{1}{1-x}$$

$$\sum_{k=1}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}$$

T = # queries to get all n

$= T_1 + T_2 + \ldots + T_n$ ← queries to get nth unique element

← waiting to get success

$T_i \sim \text{Geometric}(p_i)$ with $p_i = \frac{n-(i-1)}{n}$

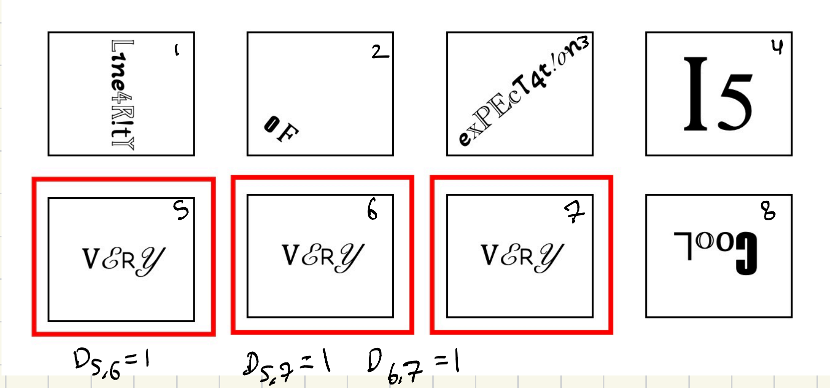
$$\mathbb{E}[T_i] = \sum_{k=1}^{\infty} (1-p_i)^{k-1} k \, p_i = p_i \frac{1}{(1-(1-p_i))^2} = \frac{1}{p_i} = \frac{n}{n-(i-1)}$$

Harmonic series

$$\mathbb{E}[T] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{n}{n-(i-1)} = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\right) = n \sum_{j=1}^{n} \frac{1}{j} \simeq n \log n$$

Define

$$D_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ same} \\ 0 & \text{else} \end{cases}$$



$$D = \sum_{i,j=1\,:\,i<j}^{m} D_{i,j}$$

$$\mathbb{E}[D] = \sum_{i<j} \mathbb{E}[D_{i,j}] = \binom{m}{2}\frac{1}{n} = \frac{m(m-1)}{2n}$$

Suppose we made $m = 1{,}000$ queries and saw $D = 10$ duplicates.

How does this compare to what we expect?

Markov's Inequality

Theorem: For any non-negative rv X and $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Proof:

$$\mathbb{E}[X] = \sum_{x} x \Pr(X = x)$$

$$= \sum_{x:\, x \geq t} x \Pr(X = x) + \sum_{x:\, x < t} x \Pr(X = x)$$

$$\geq \sum_{x:\, x \geq t} t \Pr(X = x) + 0 = t \Pr(X \geq t)$$

Answer to duplicate question:

$$\Pr(D \geq 10) \leq \frac{\mathbb{E}[D]}{10} = \frac{.4995}{10} = .04995$$