CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 2

Samson Zhou

Last Time: Class Logistics

- Course materials: https://samsonzhou.github.io/csce689-2023

- LaTeX summary of lectures 20%
- Midterm presentation 35%
- Final project 45%

Last Time: Probability Basics

- Conditional distribution: $\Pr[X = x | Y = y]$ is the probability that X achieves the value x when Y achieves the value y

$$\Pr[X = x | Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}$$

- Implies Bayes' theorem

- Random variables X and Y are independent if $\Pr[X = x] = \Pr[X = x | Y = y]$ for all possible outcomes $x \in \Omega_X, y \in \Omega_Y$

Warm-Up Question

- Suppose S_1 is a "bad" event that occurs with probability $\frac{0}{n}$
- Suppose S_2 is a "bad" event that occurs with probability $\frac{1}{n}$
- Suppose S_3 is a "bad" event that occurs with probability $\frac{2}{n}$

- What is the probability that none of the bad events occurs?

Warm-Up Question

- Suppose S_1 is a "bad" event that occurs with probability $\frac{0}{n}$
- Suppose S_2 is a "bad" event that occurs with probability $\frac{1}{n}$
- Suppose S_3 is a "bad" event that occurs with probability $\frac{2}{n}$

- What is *a lower bound* on the probability that none of the bad events occur?

Warm-Up Question

- Suppose S_1 is a "bad" event that occurs with probability $\frac{0}{n}$
- Suppose S_2 is a "bad" event that occurs with probability $\frac{1}{n}$
- Suppose S_3 is a "bad" event that occurs with probability $\frac{2}{n}$

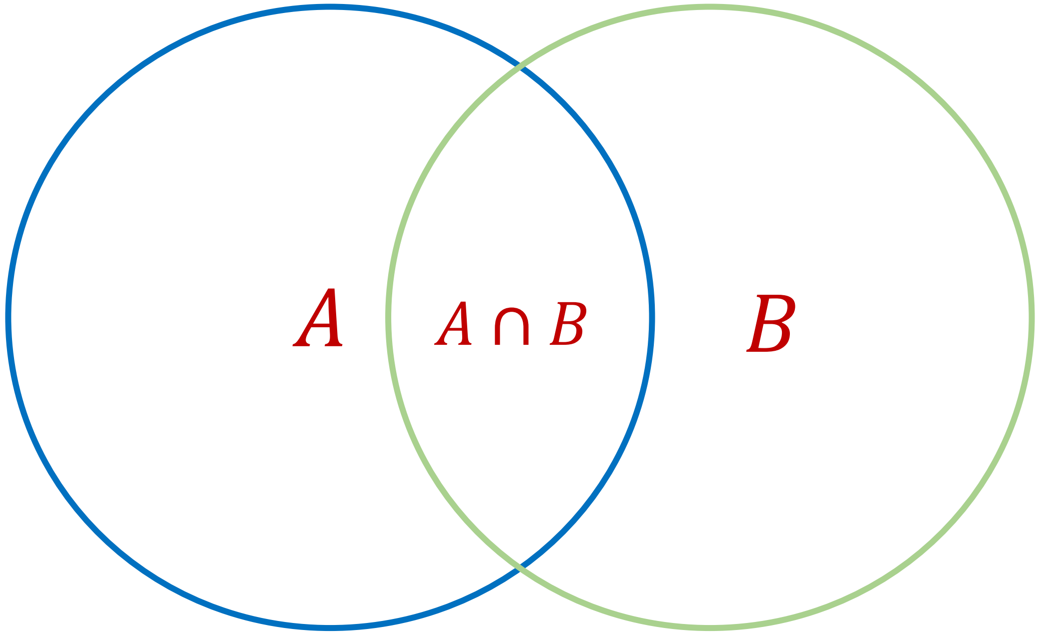
- What is *a lower bound* on the probability that none of the bad events occur? $1 - \frac{3}{n}$

Last Time: Union Bound (Boole's Inequality)

- Let $S_1,\ldots, S_k$ be a set of events that occur with probability $p_1,\ldots, p_k$
- The probability that at least one of the events $S_1,\ldots, S_k$ occurs is at most $p_1 + \cdots + p_k$
- Implication: the probability that NONE of the events $S_1,\ldots, S_k$ occur is at least $1 - (p_1 + \cdots + p_k)$

Last Time: Union Bound

- $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$



- Proof by induction

Today

- Hashing
- Abstraction: balls-in-bins
- Birthday paradox

Trivia Question #1 (Birthday Paradox)

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls? Example: 1, 5, 2, 4, 5

- $\Theta(1)$
- $\Theta(\log n)$
- $\Theta(\sqrt{n})$
- $\Theta(n)$

Trivia Question #2 (Limits)

- Let $c > 0$ be a constant. What is $\lim_{n\to\infty}\left(1 - \frac{c}{n}\right)^n$?

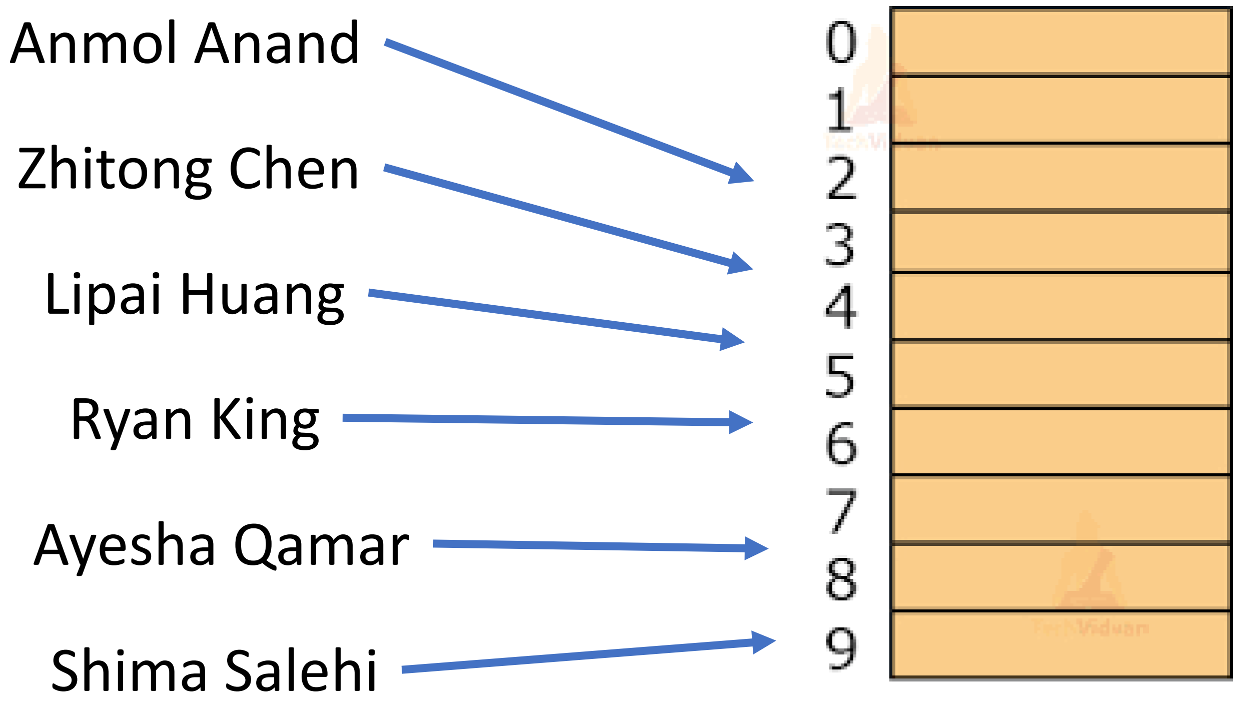
- 0
- $\frac{1}{c}$
- $\frac{1}{2c}$
- $\frac{1}{e^c}$
- 1

Hashing

- Suppose we have a number of files, how do we consistently store them in memory?

Hashing

- Suppose we have a number of files, how do we consistently store them in memory?

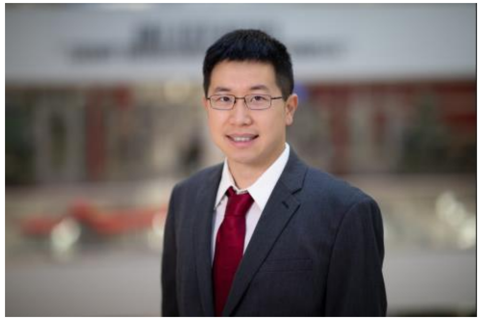


Hashing

- Suppose we have a number of files, how do we consistently store them in memory?

Hashing

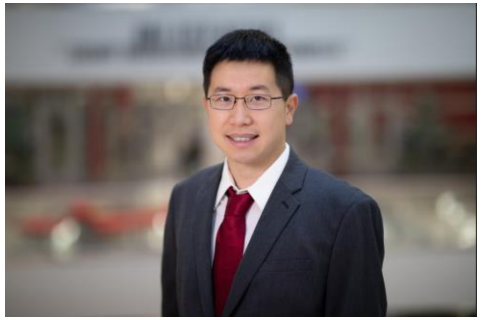
- Suppose we have a number of files, how do we consistently store them in memory?

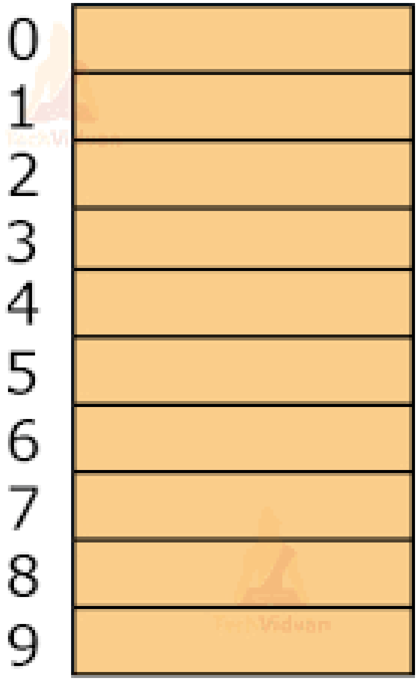


0	Anmol Anand
1	Zhitong Chen
2	Lipai Huang
3	Ryan King
4	Ayesha Qamar
5	Shima Salehi
6	
7	
8	
9	

Hashing

- Suppose we have a number of files, how do we consistently store them in memory?

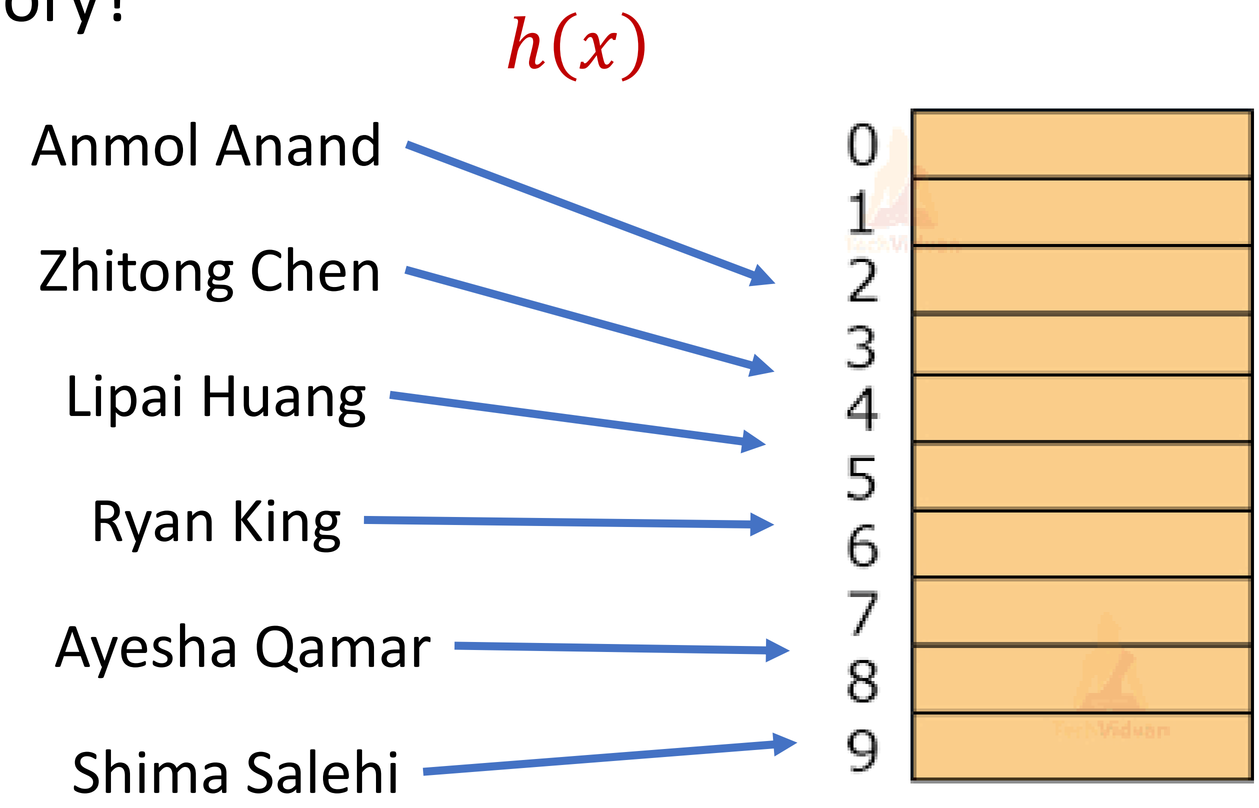




- Goal: Fast query time

Hashing

- Suppose we have a number of files, how do we consistently store them in memory?



Hash Tables

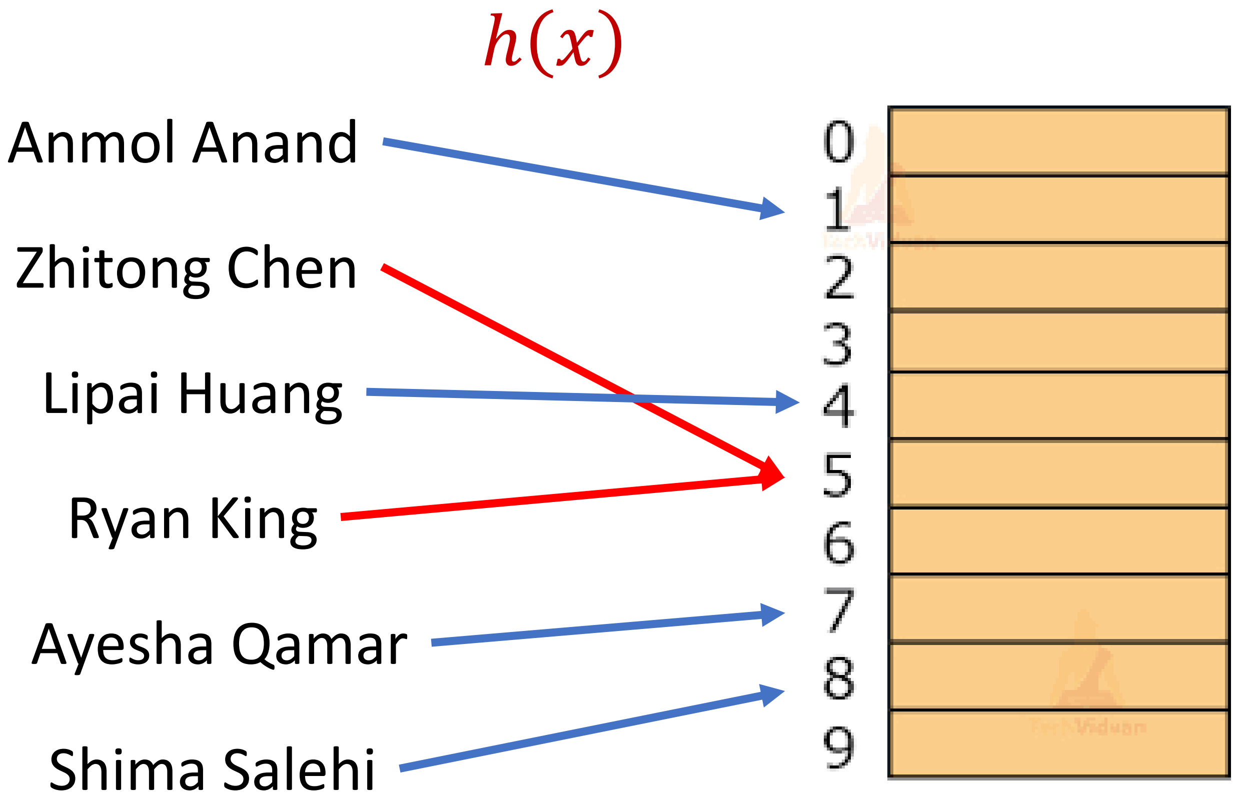
- We have a set of m items from some large universe that we want to store into a database (images, text documents, IP addresses) with n locations

- Goal: $\mathrm{query}(x)$ to check if the database contains x in $O(1)$ time

- Hash function $h: U \to [n]$ maps items from the universe to a location in the database

Collisions

- Hash function $h: U \rightarrow [n]$ maps items from the universe to a location in the database
- For $|U| \gg n$, many items map to the same location
- Collision: when multiple items should be stored in the same location

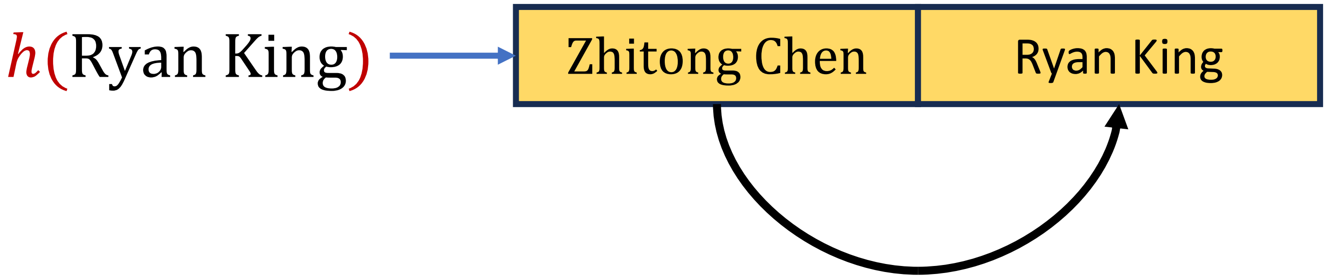


Dealing with Collisions

- Many ways of dealing with collisions
  - Store multiple items in the same location as a linked list
  - Bump item to the next available spot
  - Bump item to the next available spot using another hash function
  - Power-of-two-choices

Dealing with Collisions

- Suppose we store multiple items in the same location as a linked list



- If the maximum number of collisions in a location is c, then could traverse a linked list of size c for a query

- Query runtime: $O(c)$

Dealing with Collisions

- Goal: minimize c, the maximum number of collisions in a location
- In the worst case, all items could hash to the same location, $c = m$
- Assume the hash function h is chosen "randomly"

Random Hash Function

- Let $h: U \to [n]$ be a random hash function, so that for each $x \in U$, we have that $\Pr[h(x) = i] = \frac{1}{n}$, for all $i \in [n]$

- Assume independence, i.e., $h(x)$ and $h(y)$ are independent for any $x, y \in U$

- Suppose we insert m elements into a hash table with n locations using a random hash function. How do we analyze the number of pairwise collisions?

Birthday Paradox

- Suppose we have a room with 367 people. What is the probability that two people share the same birthday?

Birthday Paradox

- Suppose we have a room with 367 people. What is the probability that two people share the same birthday?

- Suppose we have a room with 23 people. What is the probability that two people share the same birthday?

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right)$$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4,...$ times. What is the probability we see a repeated outcome among the rolls?

$$\left(1 - \frac{0}{n}\right)\left(1 - \frac{1}{n}\right)\ldots\left(1 - \frac{k-1}{n}\right) < \frac{1}{2} \qquad \text{for} \qquad k = O(\sqrt{n})$$

Birthday Paradox

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls?
- $O(\sqrt{n})$

- But is it $\Theta(\sqrt{n})$?

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

- Let S_i be the event that the i-th roll is a repeated outcome, conditioned on the previous rolls not being a repeated outcome
- $\Pr[S_i] = \frac{i-1}{n}$
- $\Pr[S_1 \cup \cdots \cup S_k] \leq ???$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4,\ldots$ times. What is the probability we see a repeated outcome among the rolls?

- Let S_i be the event that the i-th roll is a repeated outcome, conditioned on the previous rolls not being a repeated outcome
- $\Pr[S_i] = \frac{i-1}{n}$
- $\Pr[S_1 \cup \cdots \cup S_k] \leq \frac{0}{n} + \ \ldots + \frac{k-1}{n} \leq \frac{k^2}{n}$

Birthday Paradox

- Suppose we have a fair n-sided die that we roll $k = 1, 2, 3, 4$,... times. What is the probability we see a repeated outcome among the rolls?

- Let S_i be the event that the i-th roll is a repeated outcome, conditioned on the previous rolls not being a repeated outcome
- $\Pr[S_i] = \frac{i-1}{n}$
- $\Pr[S_1 \cup \cdots \cup S_k] \leq \frac{0}{n} + \ \ldots + \frac{k-1}{n} \leq \frac{k^2}{n}$

Union Bound

Birthday Paradox

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls?
- $\Theta(\sqrt{n})$

Trivia Question #1 (Birthday Paradox)

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we see a repeated outcome among the rolls? Example: 1, 5, 2, 4, 5

- $\Theta(1)$
- $\Theta(\log n)$
- $\Theta(\sqrt{n})$
- $\Theta(n)$

Trivia Question #2 (Limits)

- Let $c > 0$ be a constant. What is $\lim_{n\to\infty}\left(1 - \frac{c}{n}\right)^n$?

- 0
- $\frac{1}{c}$
- $\frac{1}{2c}$
- $\frac{1}{e^c}$
- 1

CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 3

Samson Zhou

Trivia Question #3 (Coupon Collector)

- Suppose we have a fair n-sided die. “On average”, how many times should we roll the die before we all possible outcomes among the rolls? Example: $1, 5, 2, 4, 1, 3, 1, 6$ for $n = 6$

- $\Theta(n)$
- $\Theta(n \log n)$
- $\Theta(n\sqrt{n})$
- $\Theta(n^2)$

Trivia Question #4 (Max Load)

- Suppose we have a fair n-sided die that we roll n times. “On average”, what is the largest number of times any outcome is rolled? Example: $1, 5, 2, 4, 1, 3, 1$ for $n = 7$

- $\Theta(1)$
- $\tilde{\Theta}(\log n)$
- $\tilde{\Theta}(\sqrt{n})$
- $\tilde{\Theta}(n)$

Expected Value

- The expected value of a random variable X over Ω is:

$$\mathrm{E}[X] = \sum_{x \in \Omega} \Pr[X = x] \cdot x$$

- The "average value of the random variable"

- Linearity of expectation: $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$

Expected Value

- Suppose we roll a 6-sided die
- Let X be the outcome of the roll
- What is $\mathrm{E}[X]$?

Moments

- For $p > 0$, the p-th moment of a random variable X over Ω is:

$$\mathrm{E}[X^p] = \sum_{x \in \Omega} \Pr[X = x] \cdot x^p$$

Variance

- The variance of a random variable X over Ω is:

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

- Linearity of variance for *independent* random variables: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

- “How far numbers are from the average”

Variance

- Suppose X takes the value 1 with probability $\frac{1}{2}$ and takes the value -1 with probability $\frac{1}{2}$

- What is $\mathrm{E}[X]$?

- What is $\mathrm{Var}[X]$?

Variance

- Suppose Y takes the value 100 with probability $\frac{1}{2}$ and takes the value -100 with probability $\frac{1}{2}$

- What is $\mathrm{E}[Y]$?

- What is $\mathrm{Var}[Y]$?

Chebyshev's Inequality

- Let X be a random variable with expected value $\mu := \mathrm{E}[X]$ and variance $\sigma^2 := \mathrm{Var}[X]$

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

- "What is the probability a random variable is far away from its average?"