Example: Finding the $k$-Smallest Element in an ordered set.

Procedure Order$(S, k)$;

**Input:** A set $S$, an integer $k \leq |S| = n$.

**Output:** The $k$ smallest element in the set $S$. 

Example: Finding the $k$-Smallest Element

Procedure Order($S, k$);

**Input:** A set $S$, an integer $k \leq |S| = n$.

**Output:** The $k$ smallest element in the set $S$.

1. If $|S| = k = 1$ return $S$.
2. Choose a random element $y$ uniformly from $S$.
3. Compare all elements of $S$ to $y$. Let $S_1 = \{x \in S \mid x \leq y\}$ and $S_2 = \{x \in S \mid x > y\}$.
4. If $k \leq |S_1|$ return Order($S_1, k$) else return Order($S_2, k - |S_1|$).

**Theorem**

1. *The algorithm always returns the $k$-smallest element in $S$*
2. *The algorithm performs $O(n)$ comparisons in expectation.*
Random Variable

Definition

A random variable $X$ on a sample space $\Omega$ is a real-valued function on $\Omega$; that is, $X : \Omega \rightarrow \mathbb{R}$. A discrete random variable is a random variable that takes on only a finite or countably infinite number of values.

Discrete random variable $X$ and real value $a$: the event “$X = a$” represents the set $\{s \in \Omega : X(s) = a\}$.

$$\Pr(X = a) = \sum_{s \in \Omega : X(s) = a} \Pr(s)$$
Independence

**Definition**

Two random variables $X$ and $Y$ are independent if and only if

$$
\Pr((X = x) \cap (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y)
$$

for all values $x$ and $y$. Similarly, random variables $X_1, X_2, \ldots, X_k$ are mutually independent if and only if for any subset $I \subseteq [1, k]$ and any values $x_i, i \in I$,

$$
\Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i).
$$
Expectation

Definition

The expectation of a discrete random variable $X$, denoted by $E[X]$, is given by

$$E[X] = \sum_i i \Pr(X = i),$$

where the summation is over all values in the range of $X$. The expectation is finite if $\sum_i |i| \Pr(X = i)$ converges; otherwise, the expectation is unbounded.

The expectation (or mean or average) is a weighted sum over all possible values of the random variable.
Median

Definition

The **median** of a random variable $X$ is a value $m$ such that

$$\Pr(X < m) \leq 1/2 \quad \text{and} \quad \Pr(X > m) < 1/2.$$
# Linearity of Expectation

### Theorem

*For any two random variables $X$ and $Y*

\[
E[X + Y] = E[X] + E[Y].
\]

### Lemma

*For any constant $c$ and discrete random variable $X,*

\[
E[cX] = cE[X].
\]
Example: Finding the $k$-Smallest Element

Procedure $\text{Order}(S, k)$;

**Input:** A set $S$, an integer $k \leq |S| = n$.

**Output:** The $k$ smallest element in the set $S$.

1. If $|S| = k = 1$ return $S$.
2. Choose a random element $y$ uniformly from $S$.
3. Compare all elements of $S$ to $y$. Let $S_1 = \{x \in S \mid x \leq y\}$ and $S_2 = \{x \in S \mid x > y\}$.
4. If $k \leq |S_1|$ return $\text{Order}(S_1, k)$ else return $\text{Order}(S_2, k - |S_1|)$.

**Theorem**

1. The algorithm always returns the $k$-smallest element in $S$
2. The algorithm performs $O(n)$ comparisons in expectation.
Proof

• We say that a call to $\text{Order}(S, k)$ was successful if the random element was in the middle $1/3$ of the set $S$. A call is successful with probability $1/3$.

• After the $i$-th successful call the size of the set $S$ is bounded by $n(2/3)^i$. Thus, need at most $\log_{3/2} n$ successful calls.

• Let $X$ be the total number of comparisons. Let $T_i$ be the number of iterations between the $i$-th successful call (included) and the $i + 1$-th (excluded):

$$E[X] \leq \sum_{i=0}^{\log_{3/2} n} n(2/3)^i E[T_i].$$

• $T_i$ has a geometric distribution $G(1/3)$. 

The Geometric Distribution

Definition

A geometric random variable $X$ with parameter $p$ is given by the following probability distribution on $n = 1, 2, \ldots$.

$$Pr(X = n) = (1 - p)^{n-1}p.$$ 

Example: repeatedly draw independent Bernoulli random variables with parameter $p > 0$ until we get a 1. Let $X$ be number of trials up to and including the first 1. Then $X$ is a geometric random variable with parameter $p$. 

Lemma

Let $X$ be a discrete random variable that takes on only non-negative integer values. Then

$$E[X] = \sum_{i=1}^{\infty} \Pr(X \geq i).$$

Proof.

$$\sum_{i=1}^{\infty} \Pr(X \geq i) = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \Pr(X = j)$$

$$= \sum_{j=1}^{\infty} \sum_{i=1}^{j} \Pr(X = j)$$

$$= \sum_{j=1}^{\infty} j \Pr(X = j) = E[X].$$
For a geometric random variable $X$ with parameter $p$, 

$$\Pr(X \geq i) = \sum_{n=i}^{\infty} (1 - p)^{n-1} p = (1 - p)^{i-1}.$$ 

$$E[X] = \sum_{i=1}^{\infty} \Pr(X \geq i)$$

$$= \sum_{i=1}^{\infty} (1 - p)^{i-1}$$

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

$$= \frac{1}{p}$$
Proof

• Let $X$ be the total number of comparisons.
• Let $T_i$ be the number of iterations between the $i$-th successful call (included) and the $(i+1)$-th (excluded):

$$
\mathbb{E}[X] \leq \sum_{i=0}^{\log_{3/2} n} n(2/3)^i \mathbb{E}[T_i].
$$

• $T_i \sim G(1/3)$, therefore $\mathbb{E}[T_i] = 3$.
• Expected number of comparisons:

$$
\mathbb{E}[X] \leq \sum_{j=0}^{\log_{3/2} n} 3n \left(\frac{2}{3}\right)^j \leq 9n.
$$

Theorem

1. The algorithm always returns the $k$-smallest element in $S$
2. The algorithm performs $O(n)$ comparisons in expectation.

What is the probability space?
Finding the $k$-Smallest Element with no Randomization

Procedure Det-Order($S, k$);

Input: An array $S$, an integer $k \leq |S| = n$.

Output: The $k$ smallest element in the set $S$.

1. If $|S| = k = 1$ return $S$.
2. Let $y$ be the first element in $S$.
3. Compare all elements of $S$ to $y$. Let $S_1 = \{x \in S \mid x \leq y\}$ and $S_2 = \{x \in S \mid x > y\}$.
4. If $k \leq |S_1|$ return Det-Order($S_1, k$) else return Det-Order($S_2, k - |S_1|$).

Theorem

The algorithm returns the $k$-smallest element in $S$ and performs $O(n)$ comparisons in expectation over all possible input permutations.
Randomized Algorithms:

- Analysis is true for any input.
- The sample space is the space of random choices made by the algorithm.
- Repeated runs are independent.

Probabilistic Analysis:

- The sample space is the space of all possible inputs.
- If the algorithm is deterministic repeated runs give the same output.
A **Monte Carlo Algorithm** is a randomized algorithm that may produce an incorrect solution.

For decision problems: A **one-side error** Monte Carlo algorithm errs only one one possible output, otherwise it is a **two-side error** algorithm.

A **Las Vegas** algorithm is a randomized algorithm that **always** produces the correct output.

In both types of algorithms the run-time is a random variable.
Expectation is not everything. . .

Which algorithm do you prefer?

1. Algorithm I: takes 1 minute with probability 0.99, but with probability 0.01 takes an hour.

2. Algorithm II: takes 1 min with probability $\frac{1}{2}$ and 3 minutes with probability $\frac{1}{2}$. 
Expectation is not everything...

Which algorithm do you prefer?

1. Algorithm I: takes 1 minute with probability 0.99, but with probability 0.01 takes an hour. (Expected run-time 1.6.)

2. Algorithm II: takes 1 min with probability $\frac{1}{2}$ and 3 minutes with probability $\frac{1}{2}$. (Expected run-time 2.)

In addition to expectation we need a bound on the probability that the run time of the algorithm deviates significantly from its expectation.
Bounding Deviation from Expectation

Theorem

[Markov Inequality] For any non-negative random variable $X$, and for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{E[X]}{a}.$$  

Proof.

$$E[X] = \sum i \Pr(X = i) \geq a \sum_{i \geq a} \Pr(X = i) = a \Pr(X \geq a).$$

Example: The expected number of comparisons executed by the $k$-select algorithm was $9n$. The probability that it executes $18n$ comparisons or more is

$$\leq \frac{9n}{18n} = \frac{1}{2}.$$
Variance

Definition
The variance of a random variable $X$ is


Definition
The standard deviation of a random variable $X$ is

$$\sigma(X) = \sqrt{Var[X]}.$$
Chebyshev’s Inequality

**Theorem**

For any random variable $X$, and any $a > 0$,

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$ 

**Proof.**

$$Pr(|X - E[X]| \geq a) = Pr((X - E[X])^2 \geq a^2)$$

By Markov inequality

$$Pr((X - E[X])^2 \geq a^2) \leq \frac{E[(X - E[X])^2]}{a^2}$$

$$= \frac{Var[X]}{a^2}$$
Theorem

For any random variable $X$ and any $a > 0$:

$$Pr(|X - E[X]| \geq a\sigma[X]) \leq \frac{1}{a^2}.$$

Theorem

For any random variable $X$ and any $\varepsilon > 0$:

$$Pr(|X - E[X]| \geq \varepsilon E[X]) \leq \frac{\text{Var}[X]}{\varepsilon^2 (E[X])^2}.$$
Theorem

If $X$ and $Y$ are independent random variables

$$E[XY] = E[X] \cdot E[Y].$$

Proof.

$$E[XY] = \sum_{i} \sum_{j} i \cdot j \Pr((X = i) \cap (Y = j)) =$$

$$\sum_{i} \sum_{j} ij \Pr(X = i) \cdot \Pr(Y = j) =$$

$$\left( \sum_{i} i \Pr(X = i) \right) \left( \sum_{j} j \Pr(Y = j) \right).$$
Theorem

If $X$ and $Y$ are independent random variables

\[ \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]. \]

Proof.

\[
\text{Var}[X + Y] = E[(X + Y - E[X] - E[Y])^2] = \\
E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] = \\
\text{Var}[X] + \text{Var}[Y] + 2E[X - E[X]]E[Y - E[Y]]
\]

Since the random variables $X - E[X]$ and $Y - E[Y]$ are independent.

But $E[X - E[X]] = E[X] - E[X] = 0$. \qed
Let $X$ be a 0-1 random variable such that

$$Pr(X = 1) = p, \quad Pr(X = 0) = 1 - p.$$  

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p.$$  

A Binomial Random variable

Consider a sequence of $n$ independent Bernoulli trials $X_1, \ldots, X_n$. Let

$$X = \sum_{i=1}^{n} X_i.$$  

$X$ has a **Binomial** distribution $X \sim B(n, p)$.

$$Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$  

$$E[X] = np.$$  

$$Var[X] = np(1 - p).$$
The Geometric Distribution

• How many times do we need to perform a trial with probability $p$ for success till we get the first success?
• How many times do we need to roll a dice until we get the first 6?

Definition

A geometric random variable $X$ with parameter $p$ is given by the following probability distribution on $n = 1, 2, \ldots$

$$\Pr(X = n) = (1 - p)^{n-1} p.$$
For a geometric random variable with parameter $p$ and $n > 0$,

$$\Pr(X = n + k \mid X > k) = \Pr(X = n).$$

Proof.

$$\Pr(X = n + k \mid X > k) = \frac{\Pr((X = n + k) \cap (X > k))}{\Pr(X > k)}$$

$$= \frac{\Pr(X = n + k)}{\Pr(X > k)} = \frac{(1 - p)^{n+k-1}p}{\sum_{i=k}^{\infty}(1 - p)^i p}$$

$$= \frac{(1 - p)^{n+k-1}p}{(1 - p)^k} = (1 - p)^{n-1}p = \Pr(X = n).$$
**Conditional Expectation**

**Definition**

\[
E[Y \mid Z = z] = \sum_y y \Pr(Y = y \mid Z = z),
\]

where the summation is over all \( y \) in the range of \( Y \).
Lemma

For any random variables $X$ and $Y$,

$$E[X] = E_y[E_X[X \mid Y]] = \sum_y \Pr(Y = y)E[X \mid Y = y],$$

where the sum is over all values in the range of $Y$.

Proof.

\[
\sum_y \Pr(Y = y)E[X \mid Y = y] \\
= \sum_y \Pr(Y = y) \sum_x x \Pr(X = x \mid Y = y) \\
= \sum_x \sum_y x \Pr(X = x \mid Y = y) \Pr(Y = y) \\
= \sum_x \sum_y x \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = E[X].
\]
Example

Consider a two phase game:

- Phase I: roll one die. Let $X$ be the outcome.
- Phase II: Flip $X$ fair coins, let $Y$ be the number of HEADs.
- You receive a dollar for each HEAD.

$Y$ is distributed $B(X, \frac{1}{2})$,

$$E[Y \mid X = a] = \frac{a}{2}$$

$$E[Y] = \sum_{i=1}^{6} E[Y \mid X = i]Pr(X = i)$$

$$= \sum_{i=1}^{6} \frac{i}{2}Pr(X = i) = \frac{7}{4}$$
Geometric Random Variable: Expectation

- Let $X$ be a geometric random variable with parameter $p$.
- Let $Y = 1$ if the first trial is a success, $Y = 0$ otherwise.

$$
E[X] = \Pr(Y = 0)E[X \mid Y = 0] + \Pr(Y = 1)E[X \mid Y = 1]
= (1 - p)E[X \mid Y = 0] + pE[X \mid Y = 1].
$$

- If $Y = 0$ let $Z$ be the number of trials after the first one.

Variance of a Geometric Random Variable

- We use

\[\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\]

- To compute \(\mathbb{E}[X^2]\), let \(Y = 1\) if the first trial is a success, \(Y = 0\) otherwise.

\[
\mathbb{E}[X^2] = \Pr(Y = 0)\mathbb{E}[X^2 \mid Y = 0] + \Pr(Y = 1)\mathbb{E}[X^2 \mid Y = 1]
= (1 - p)\mathbb{E}[X^2 \mid Y = 0] + p\mathbb{E}[X^2 \mid Y = 1].
\]

- If \(Y = 0\) let \(Z\) be the number of trials after the first one.

\[
\mathbb{E}[X^2] = (1 - p)\mathbb{E}[(Z + 1)^2] + p \cdot 1
= (1 - p)\mathbb{E}[Z^2] + 2(1 - p)\mathbb{E}[Z] + 1,
\]
• $\mathbb{E}[Z] = 1/p$ and $\mathbb{E}[Z^2] = \mathbb{E}[X^2]$. 

•

\[
\mathbb{E}[X^2] = (1 - p)\mathbb{E}[(Z + 1)^2] + p \cdot 1 \\
= (1 - p)\mathbb{E}[Z^2] + 2(1 - p)\mathbb{E}[Z] + 1,
\]

•

\[
\mathbb{E}[X^2] = (1 - p)\mathbb{E}[X^2] + 2(1 - p)/p + 1 = (1 - p)\mathbb{E}[X^2] + (2 - p)/p,
\]

• $\mathbb{E}[X^2] = (2 - p)/p^2$. 

Variance of a Geometric Random Variable

\[
\text{Var}[X] = \mathsf{E}[X^2] - \mathsf{E}[X]^2
\]

\[
= \frac{2 - p}{p^2} - \frac{1}{p^2}
\]

\[
= \frac{1 - p}{p^2}.
\]
Back to the $k$-select Algorithm

- Let $X$ be the total number of comparisons.
- Let $T_i$ be the number of iterations between the $i$-th successful call (included) and the $i + 1$-th (excluded):
  - $X \leq \sum_{i=0}^{\log_3 2} n (2/3)^i T_i$.
  - $T_i \sim G(1/3)$, therefore $E[T_i] = 3$, $Var[T_i] = 9/4$.
- Expected number of comparisons:
  $E[X] \leq \sum_{j=0}^{\log_3 2} 3n (2/3)^j \leq 9n$.
- Variance of the number of comparisons:
  $Var[X] = \sum_{i=0}^{\log_3 2} n^2 (2/3)^{2i} Var[T_i] \leq 11n^2$

$$\Pr(|X - E[X]| \geq \delta E[X]) \leq \frac{Var[X]}{\delta^2 E[X]^2} \leq \frac{11n^2}{\delta^2 36n^2}$$
Example: Coupon Collector’s Problem

Suppose that each box of cereal contains a random coupon from a set of $n$ different coupons. How many boxes of cereal do you need to buy before you obtain at least one of every type of coupon?

Let $X$ be the number of boxes bought until at least one of every type of coupon is obtained. Let $X_i$ be the number of boxes bought while you had exactly $i - 1$ different coupons.

\[ X = \sum_{i=1}^{n} X_i \]

$X_i$ is a geometric random variable with parameter

\[ p_i = 1 - \frac{i - 1}{n}. \]
\[ \mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n - i + 1}. \]

\[ \mathbb{E}[X] = E \left[ \sum_{i=1}^{n} X_i \right] \]

\[ = \sum_{i=1}^{n} \mathbb{E}[X_i] \]

\[ = \sum_{i=1}^{n} \frac{n}{n - i + 1} \]

\[ = n \sum_{i=1}^{n} \frac{1}{i} = n \ln n + \Theta(n). \]
Example: Coupon Collector’s Problem

- We place balls independently and uniformly at random in $n$ boxes.
- Let $X$ be the number of balls placed until all boxes are not empty.
- What is $E[X]$?
• Let $X_i =$ number of balls placed when there were exactly $i - 1$ non-empty boxes.

• $X = \sum_{i=1}^{n} X_i$.

• $X_i$ is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.

•

$$E[X_i] = \frac{1}{p_i} = \frac{n}{n - i + 1}.$$  

$$E[X] = E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n - i + 1} = n \ln n + \Theta(n).$$
Back to the Coupon Collector’s Problem

• Suppose that each box of cereal contains a random coupon from a set of $n$ different coupons.
• Let $X$ be the number of boxes bought until at least one of every type of coupon is obtained.
• $E[X] = nH_n = n \ln n + \Theta(n)$
• What is $\Pr(X \geq 2E[X])$?
• Applying Markov’s inequality

$$\Pr(X \geq 2nH_n) \leq \frac{1}{2}.$$  

• Can we do better?
Let $X_i$ be the number of boxes bought while you had exactly $i - 1$ different coupons.

$X = \sum_{i=1}^{n} X_i$.

$X_i$ is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.

$\text{Var}[X_i] \leq \frac{1}{p^2} \leq \left(\frac{n}{n-i+1}\right)^2$.

\[
\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] \leq \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 \leq \frac{\pi^2 n^2}{6}.
\]

By Chebyshev’s inequality

\[
\Pr(|X - nH_n| \geq nH_n) \leq \frac{n^2 \pi^2 / 6}{(nH_n)^2} = \frac{\pi^2}{6(nH_n)^2} = O\left(\frac{1}{\ln^2 n}\right).
\]
The probability of not obtaining the $i$-th coupon after $n \ln n + cn$ steps:

\[
\left(1 - \frac{1}{n}\right)^{n(\ln n + c)} \leq e^{-(\ln n + c)} = \frac{1}{e^c n}.
\]

By a union bound, the probability that some coupon has not been collected after $n \ln n + cn$ step is $e^{-c}$.

The probability that all coupons are not collected after $2n \ln n$ steps is at most $1/n$. 