Expectation is not everything...

With probability 0.99, you owe me $\$10$; with probability 0.01, I owe you $\$1000$. Would you play this game?

Expectation is not everything. . .

With probability 0.99, you owe me \$10; with probability 0.01, I owe you \$1000. Would you play this game?

X = *your* winning (in \$). $\mathbf{E}[X] = 0.1$.

Y = *my* winning (in \$). $\mathbf{E}[Y] = -0.1$.

Expectation is not everything...

With probability 0.99, you owe me \$10; with probability 0.01, I owe you \$1000. Would you play this game?

X = *your* winning (in \$). $\mathbf{E}[X] = 0.1$. But $Pr(X \geq \mathbf{E}[X]) = 0.01$ (in fact, $Pr(X \geq 0) = 0.01$)

Y = *my* winning (in \$). $\mathbf{E}[Y] = -0.1$. But $Pr(Y \geq \mathbf{E}[Y]) = 0.99$ (in fact, $Pr(Y > 0) = 0.99$)

Which algorithm would you prefer?

1. The expected run time of the algorithm is 1 hour but it will take 100 hours in 1% of the runs.
2. The run time is always 2 hours.

We need to bound the probability that the run time of the algorithm deviates significantly from its average.

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.

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

$\square$

Example: What is the probability of getting more than $\frac{3N}{4}$ heads in N coin flips? $\leq \frac{N/2}{3N/4} \leq \frac{2}{3}$.

Variance

Definition

The **variance** of a random variable X is

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

Definition

The **standard deviation** of a random variable X is

$$\sigma(X) = \sqrt{Var[X]}.$$

Example: Let X be a 0-1 random variable with $Pr(X=0)=Pr(X=1)=1/2$.

$$E[X]=1/2.$$

$$\begin{aligned} Var[X] &= E[(X-E[X])^2] \\ &= \frac{1}{2}\left(1-\frac{1}{2}\right)^2+\frac{1}{2}\left(0-\frac{1}{2}\right)^2 \\ &= \frac{1}{4} \end{aligned}$$

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* $\varepsilon > 0$*:*

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

Theorem

If X and Y are independent random variables

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

Proof.

$$E[XY] = \sum_i \sum_j i \cdot j Pr((X = i) \cap (Y = j)) \; =$$

$$\sum_i \sum_j ij Pr(X = i) \cdot Pr(Y = j) =$$

$$\left(\sum_i i Pr(X = i) \right) \left(\sum_j j Pr(Y = j) \right).$$

$\square$

Theorem

If X and Y are independent random variables

$$Var[X+Y] = Var[X] \; + \; Var[Y].$$

Proof.

$$Var[X+Y] = E[(X+Y-E[X]-E[Y])^2] =$$

$$E[(X-E[X])^2 + (Y-E[Y])^2 + 2(X-E[X])(Y-E[Y])] =$$

$$Var[X] + Var[Y] + 2E[X-E[X]]E[Y-E[Y]]$$

Since the random variables $X - E[X]$ and $Y - E[Y]$ are independent.
But $E[X-E[X]] = E[X] - E[X] = 0.$ $\square$

Bernoulli Trial

Let X be a 0-1 random variable such that

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

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

$$Var[X] = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)(1-p+p) =$$

$$p(1-p).$$

A Binomial Random variable

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

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

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

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

$$E[X] = np.$$

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

Back to Coin Flips

Assume again that we flip N coins. Let X be the number of heads.
$X_i = 1$ if the i-th flip was a head else $X_i = 0$.
$E[X_i] = 1/2$. $Var[X_i] = 1/4$.

$$Pr(X \geq 3N/4) \leq Pr(|X - E[X]| \geq N/4) =$$

$$Pr(|X - E[X]| \geq E[X]/2) \leq \frac{Var[X]}{(E[X])^2(1/4)} =$$

$$\frac{N/4}{(N^2/4)(1/4)} = 4/N.$$

A significantly better bound than $2/3$.

The Advantage of Multiple Samples

Theorem

For any random variable X and constant a,

$$Var[aX] = a^2 Var[X].$$

Proof.

$$\begin{aligned} Var[aX] &= E[(aX - E[aX])^2] = E[a^2(X - E[X])^2] \\ &= a^2 E[(X - E[X])^2] = a^2 Var[X]. \end{aligned}$$

$\square$

Theorem

Let $X_1, \ldots, X_n$ *be* n *independent, identically distributed random variables. Let* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} Var[X_i].$$

The (Weak) Law of Large Numbers

Theorem

Let $X_1, \ldots, X_n$ be independent, identically distributed, random variables. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. For any constant $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|\bar{X}_n - \mathbf{E}[X_i]| \le \varepsilon) = 1.$$

Proof.

$Var[\bar{X}_n] = \frac{1}{n} Var[X_i]$. Applying Chebyshev's inequality

$$\Pr(|\bar{X}_n - \mathbf{E}[X_i]| > \varepsilon) \le \frac{Var[X_i]}{n\varepsilon^2}.$$

□

[Can be proven even when $Var[X_i]$ is not bounded.]

Algorithm for Computing the Median

The **median** of a set S of n distinct elements is the $\lceil \frac{n}{2} \rceil$ largest element in the set.
If $n = 2k + 1$, the median element is the $k + 1$-th element in the sorted order.

Algorithm for Computing the Median

The **median** of a set S of n distinct elements is the $\lceil \frac{n}{2} \rceil$ largest element in the set.

If $n = 2k + 1$, the median element is the $k + 1$-th element in the sorted order.

Easily computed through sorting in $O(n \log n)$ time. There exists a complicated $O(n)$ deterministic algorithm.

Randomized Median Algorithm

Input: A set S of $n = 2k + 1$ elements from a totally ordered universe.

Output: The $k + 1$-th largest element in the set.

1. Pick a (multi)-set R of $s = n^{3/4}$ elements in S, chosen independently and uniformly at random with replacement. Sort the set R.
2. Let d be the $(\frac{1}{2}n^{3/4} - \sqrt{n})$-th smallest element in the sorted set R.
3. Let u be the $(\frac{1}{2}n^{3/4} + \sqrt{n})$-th smallest element in the sorted set R.
4. By comparing every element in S to d and u compute the set $C = \{x \in S \ : \ d \leq x \leq u\}$, and the numbers $\ell_d = |\{x \in S \ : \ x < d\}|$ and $\ell_u = |\{x \in S \ : \ x > u\}|$.
5. If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL.
6. If $|C| \leq 4n^{3/4}$ then sort the set C, otherwise FAIL.
7. Output the $(\lfloor \frac{n}{2} \rfloor - \ell_d + 1)$-st element in the sorted order of C.

Analysis

Theorem

The randomized median algorithm terminates in $O(n)$ time.

Theorem

If the randomized median algorithm does not output FAIL, then it outputs the median of S.

Analysis

Theorem

The randomized median algorithm succeeds with probability $\geq 1 - \frac{1}{n^{1/4}}$.

Intuition

- We can sort sets of size $< n/\log n$ in linear time.
- The sample of R elements are spaced "more or less" evenly among the elements of S.
- W.h.p. more than $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples are smaller than the median.
- W.h.p. more than $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples are larger than the median.
- W.h.p. the median is in the set C, and $|C| \leq 4n^{3/4}$.

Let Y_1 be the number of samples below the median.
Let Y_2 be the number of samples above the median.
The algorithm fails to compute the median in $O(n)$ time if and only if at least one of the following three events occurs:

1. $E_1: \; Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}$.
2. $E_2: \; Y_2 < \frac{1}{2}n^{3/4} - \sqrt{n}$.
3. $E_3: \; |C| > 4n^{3/4}$.

What is the probability that the three random variables Y_1, Y_2 and $|C|$ are all within the required ranges?.

The sample space in execution of this algorithm is the set of all possible choices of $n^{3/4}$ elements from n, with repetitions. (The sample space has $\Theta(n^{n^{3/4}})$ points.)
Each point in the sample space defines values for Y_1, Y_2 and $|C|$.
Computing the probabilities directly is too complicated, instead we use bounds on deviation from the expectation.

Y_1 be the number of samples below the median.
What is the probability that $Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}$?
Viewing Y_1 as the sum of $n^{3/4}$ independent 0-1 random variables, each with expectation $1/2$ and variance $1/4$ we prove (not counting the median itself):

$$E[Y_1] = \frac{1}{2}n^{3/4}.$$

$$Var[Y_1] = \frac{1}{4}n^{3/4}.$$

Applying Chebyshev Inequality we get:

$$Pr(E_1) = Pr\left(Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}\right) \leq Pr(|Y_1 - E[Y_1]| > \sqrt{n}) \leq$$

$$\frac{Var[Y_1]}{n} = \frac{n^{3/4}/4}{n} = \frac{1}{4}n^{-1/4}.$$

Similarly

$$Pr(E_2) = Pr\left(Y_2 < \frac{1}{2}n^{3/4} - \sqrt{n}\right) \leq \frac{1}{4}n^{-1/4}.$$

$$Pr(E_1 \cup E_2) \leq \frac{1}{2}n^{-1/4}.$$

Recall: $E_3 : \ |C| > 4n^{3/4}$.

Lemma

$$\Pr(E_3) \leq \frac{1}{2} n^{-1/4}.$$

Define the following two events:

1. $\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median;
2. $\mathcal{E}_{3,2}$: at least $2n^{3/4}$ elements of C are smaller than the median.

If $|C| > 4n^{3/4}$, then at least one of the above two events occurs.

We bound $\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median;

At least $2n^{3/4}$ elements of C above the median $\Rightarrow$

u is at least the $\frac{1}{2}n + 2n^{3/4}$ largest in S $\Rightarrow$

R had at least $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S.

Let X be the number of samples (in R) among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S. Let $X = \sum_{i=1}^{n^{3/4}} X_i$ where

$$X_i = \begin{cases} 1 & \text{the } i\text{-th sample in } \frac{1}{2}n - 2n^{3/4} \\ & \text{largest elements in } S \\ 0 & \text{otherwise.} \end{cases}$$

$$E[X_i] = E[(X_i)^2] = \frac{1}{2} - 2n^{-1/4}$$

$$Var[X_i] = E[(X_i)^2] - (E[X_i])^2 \le \frac{1}{4}.$$

$$E[X] = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$

$$Var[X] \le \frac{1}{4}n^{3/4}$$

Applying Chebyshev's Inequality yields

$$\begin{aligned}
\Pr(\mathcal{E}_{3,1}) &= \Pr\left(X \ge \frac{1}{2}n^{3/4} - \sqrt{n}\right) \\
&\le \Pr(|X - E[X]| \ge \sqrt{n}) \\
&\le \frac{Var[X]}{n} \le \frac{\frac{n^{\frac{3}{4}}}{4}}{n} = \frac{1}{4}n^{-\frac{1}{4}}.
\end{aligned}$$

Similarly,

$$\Pr(\mathcal{E}_{3,2}) \leq \frac{1}{4} n^{-\frac{1}{4}},$$

and

$$\Pr(E_3) \leq \Pr(\mathcal{E}_{3,1}) + \Pr(\mathcal{E}_{3,2}) \leq \frac{1}{2} n^{-\frac{1}{4}}.$$

The probability that the algorithm succeeds is

$$\geq 1 - (Pr(E_1) + Pr(E_2) + Pr(E_3)) \geq 1 - \frac{1}{n^{1/4}}.$$

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 $n = 1, 2, \ldots$.

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

Memoryless Distribution

Lemma

For a geometric random variable with parameter p and $n > 0$,

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

Proof.

$$\begin{aligned}
\Pr(X = n + k \quad \mid \quad X > k) &= \frac{\Pr((X = n + k) \cap (X > k))}{\Pr(X > k)} \\
&= \frac{\Pr(X = n + k)}{\Pr(X > k)} = \frac{(1-p)^{n+k-1}p}{\sum_{i=k}^{\infty}(1-p)^i p} \\
&= \frac{(1-p)^{n+k-1}p}{(1-p)^k} = (1-p)^{n-1}p = \Pr(X = n).
\end{aligned}$$

$\square$

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

$$p_i = 1 - \frac{i-1}{n}.$$

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

$$\begin{aligned}
\mathbf{E}[X] &= E\left[\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} \mathbf{E}[X_i] \\
&= \sum_{i=1}^{n} \frac{n}{n-i+1} \\
&= n\sum_{i=1}^{n} \frac{1}{i} = n\ln n + \Theta(n).
\end{aligned}$$

Variance of a Geometric Random Variable

- We use
$$Var[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- To compute $\mathbf{E}[X^2]$, let $Y = 1$ if the first trial is a success, $Y = 0$ otherwise.
-
$$\begin{aligned}\mathbf{E}[X^2] &= \Pr(Y = 0)\mathbf{E}[X^2 \mid Y = 0] + \Pr(Y = 1)\mathbf{E}[X^2 \mid Y = 1] \\ &= (1 - p)\mathbf{E}[X^2 \mid Y = 0] + p\mathbf{E}[X^2 \mid Y = 1].\end{aligned}$$

- If $Y = 0$ let Z be the number of trials after the first one.
-
$$\begin{aligned}\mathbf{E}[X^2] &= (1 - p)\mathbf{E}[(Z + 1)^2] + p \cdot 1 \\ &= (1 - p)\mathbf{E}[Z^2] + 2(1 - p)\mathbf{E}[Z] + 1,\end{aligned}$$

- $\mathbf{E}[Z] = 1/p$ and $\mathbf{E}[Z^2] = \mathbf{E}[X^2]$.
-
$$\begin{aligned}\mathbf{E}[X^2] &= (1-p)\mathbf{E}[(Z+1)^2] + p \cdot 1 \\ &= (1-p)\mathbf{E}[Z^2] + 2(1-p)\mathbf{E}[Z] + 1,\end{aligned}$$
-
$$\mathbf{E}[X^2] = (1-p)\mathbf{E}[X^2] + 2(1-p)/p + 1 = (1-p)\mathbf{E}[X^2] + (2-p)/p,$$
- $\mathbf{E}[X^2] = (2-p)/p^2$.

$$
\begin{aligned}
Var[X] &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 \\
&= \frac{2-p}{p^2} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}.
\end{aligned}
$$

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 $\Pr(X \geq 2\mathbf{E}[X])$?
- 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 $p_i = 1 - \frac{i-1}{n}$.
- $Var[X_i] \leq \frac{1}{p^2} \leq (\frac{n}{n-i+1})^2$.
-
$$Var[X] = \sum_{i=1}^{n} Var[X_i] \leq \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 \leq \frac{\pi^2 n^2}{6}.$$

- By Chebyshev's inequality

$$\Pr(|X - nH_n| \geq nH_n) \leq \frac{n^2\pi^2/6}{(nH_n)^2} = \frac{\pi^2}{6(H_n)^2} = O\left(\frac{1}{\ln^2 n}\right).$$

Direct Bound

- The probability of not obtaining the i-th coupon after $n \ln n + cn$ steps:

$$\left(1 - \frac{1}{n}\right)^{n(\ln n + c)} \le \mathrm{e}^{-(\ln n + c)} = \frac{1}{\mathrm{e}^{c} n}.$$

- 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 $1/n$.