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.

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

□

Example: The expected number of comparisons executed by the k-select algorithm was $9n$. The probability that it executes $18n$ comparisons or more $\leq \frac{9n}{18n} = \frac{1}{2}$.

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

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

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.

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

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

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

□

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 $E[X-E[X]] = E[X]-E[X] = 0$. □

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 $X \sim B(n, p)$.

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

$$E[X] = np.$$

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

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.

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

□

Conditional Expectation

Definition

$$\mathbf{E}[Y \mid Z = z] = \sum_{y} y \Pr(Y = y \mid Z = z),$$

where the summation is over all y in the range of Y.

Lemma

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

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

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

Proof.

$$\begin{aligned}
& \sum_y \Pr(Y = y) E[X \mid Y = y] \\
= & \sum_y \Pr(Y = y) \sum_x x \Pr(X = x \mid Y = y) \\
= & \sum_x \sum_y x \Pr(X = x \mid Y = y) \Pr(Y = y) \\
= & \sum_x \sum_y x \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = \mathbf{E}[X].
\end{aligned}$$

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 trail is a success, $Y = 0$ otherwise.
-

$$\begin{aligned}\mathbf{E}[X] &= \Pr(Y=0)\mathbf{E}[X \mid Y=0] + \Pr(Y=1)\mathbf{E}[X \mid Y=1] \\ &= (1-p)\mathbf{E}[X \mid Y=0] + p\mathbf{E}[X \mid Y=1].\end{aligned}$$

- If $Y = 0$ let Z be the number of trials after the first one.
- $\mathbf{E}[X] = (1-p)\mathbf{E}[Z+1] + p \cdot 1 = (1-p)\mathbf{E}[Z] + 1$
- But $\mathbf{E}[Z] = \mathbf{E}[X]$, giving $\mathbf{E}[X] = 1/p$.

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 $\mathbf{E}[Z^2] = \mathbf{E}[X^2]$.
-

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

-

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

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

Variance of a Geometric Random Variable

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

Back to the 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):
- $X \leq \sum_{i=0}^{\log_{3/2} n} n(2/3)^i T_i$.
- $T_i \sim G(1/3)$, therefore $\mathbf{E}[T_i] = 3$, $Var[T_i] = 9/4$.
- Expected number of comparisons:
  $\mathbf{E}[X] \leq \sum_{j=0}^{\log_{3/2} n} 3n\,(2/3)^j \leq 9n.$
- Variance of the number of comparisons:
  $Var[X] = \sum_{i=0}^{\log_{3/2} n} n^2(2/3)^{2i} Var[T_i] \leq 11n^2$

$$Pr(|X - E[X]| \geq \delta E[X]) \leq \frac{Var[X]}{\delta^2 E[X]^2} \leq \frac{11n^2}{\delta^2 81 n^2}$$

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

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

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

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

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 $\mathbf{E}[X]$?

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

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

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

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:

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

- 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 $1/n$.