CS 70 Discrete Mathematics and Probability Theory

Spring 2016 Rao and Walrand Discussion 11B Sol

1. Variance

For a random variable, X, variance of X is $E[X^2]-(E[X])^2$, calculate the variance for the following random variables

a) Calculate the variance of a random variable, X, where X represents the value of a standard 6-sided dice.

$$\Pr[X=i]=\frac{1}{6} \quad \text{for } i=1,2,\ldots,6$$

$$E[X]=\sum_{i=1}^{6} i\Pr[X=i]=\frac{1}{6}(1+2+3+\cdots+6)=\frac{7}{2}$$

$$E[X^2]=\sum_{i=1}^{6} i^2\Pr[X=i]=\frac{1}{6}(1+4+9+\cdots+36)=\frac{91}{6}$$

$$Var[X]=E[X^2]-(E[X])^2=\frac{35}{12}$$

b) Calculate the variance of a random variable defined by the binomial distribution.

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

$$X_i=\begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1-p) \end{cases}$$

and all these X_i's are i.i.d. Bernoulli r.v. So, $Var[X]=\sum_{i=1}^{n} Var[X_i]$.

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

$$E[X_i^2]=1^2\cdot p+0^2\cdot(1-p)=p$$

$$Var[X_i]=E[X_i^2]-(E[X_i])^2=p-p^2=p(1-p)$$

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

2. Geometric distribution

a) If $X_i \sim Geom(p)$, show that $\mathbb{E}[X_i^2] = \sum_{k=1}^{\infty} k(k+1)p(1-p)^{k-1} - \sum_{k=1}^{\infty} kp(1-p)^{k-1}$
By definition $\mathbb{E}[X_i^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1}$. Since $k^2 = k(k+1) - k$, we can plug that in to get the answer.

b) Use your lemma and the fact that $\mathbb{E}[X_i] = 1/p$ to simplify part a) to show $\mathbb{E}[X_i^2] = \frac{2}{p^2} - \frac{1}{p}$
For the first term, we have

$$\begin{aligned}
\sum_{k=0}^{\infty}(1-p)^k &= \frac{1}{p} && \text{take derivative of both sides twice}\\
-\sum_{k=0}^{\infty} k(1-p)^{k-1} &= -\frac{1}{p^2} && \text{1st}\\
\sum_{k=0}^{\infty} k(k-1)(1-p)^{k-2} &= \frac{2}{p^3} && \text{2nd, LHS } k=0 \text{ term is } 0\\
\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} &= \frac{2}{p^3} && \\
\sum_{j=0}^{\infty}(j+1)j(1-p)^{j-1} &= \frac{2}{p^3} && \text{change summation bounds} j = k-1\\
\sum_{j=1}^{\infty} j(j+1)(1-p)^{j-1} &= \frac{2}{p^3} && \text{LHS } j=0 \text{ term is } 0
\end{aligned}$$

We also notice that the second sum term is just the expectation of a geometric variable with parameter p. So, we get

$$\mathbb{E}[X_i^2] = p \cdot \frac{2}{p^3} - \frac{1}{p}$$

3. Chopping up DNA

In a certain biological experiment, a piece of DNA consisting of a linear sequence (or string) of 4000 nucleotides is subjected to bombardment by various enzymes. The effect of the bombardment is to randomly cut the string between pairs of adjacent nucleotides: each of the 3999 possible cuts occurs independently and with probability $1/500$.

1. What is the expected number of pieces into which the string is cut?
   Let X_i be the indicator variable for the event that a cut occurs at position i. Then, the number of pieces into which the string is cut is $X = 1 + \sum_{i=1}^{3999} X_i$ (the number of pieces is one greater than the number of cuts). By linearity of expectation,
$$\mathrm{E}[X] = 1 + \sum_{i=1}^{3999} \frac{1}{500} = 1 + 3999/500.$$

2. What is the variance of the above quantity?
   Note that
$$\mathrm{Var}[X] = \mathrm{Var}\left[1 + \sum_{i=1}^{3999} X_i\right] = \mathrm{Var}[1] + \mathrm{Var}\left[\sum_{i=1}^{3999} X_i\right] = \mathrm{Var}\left[\sum_{i=1}^{3999} X_i\right].$$
   Note that $Z = \sum_{i=1}^{3999} X_i$ is a Binomial random variable with parameters $n = 3999$ and $p = 1/500$. Hence, the variance is $np(1-p) = 3999 \cdot \frac{1}{500} \cdot \frac{499}{500}$.

3. Suppose that the cuts are no longer independent, but highly correlated: when a cut occurs in a particular location, nearby locations are much more likely to be cut as well. The probability of each individual cut remains $1/500$. Does the expected number of pieces increase, decrease, or stay the same?
   Since $\mathrm{E}[X_i]$ is still $1/500$ for each i, the expectation $\mathrm{E}[X] = 1 + \sum_{i=1}^{3999} \mathrm{E}[X_i]$ will stay the same.

4. How Many Coupons?

Consider the coupon collecting problem covered in the note. There are n distinct types of coupons that we wish to collect. Every time we buy a box, there is one coupon in it, with equal likelihood of being any one of the types of coupons. We want to figure out how many boxes we need to buy in order to get one of each coupon. For this problem, we want to bound the probability that we have to buy lots of coupons — say substantially more than $n \ln n$ coupons.

a. We represent X, the number of boxes we have to buy, as a sum of other random variables. Let X_i represent the number of boxes you buy to go from $i-1$ to i distinct coupons in your hand. The let $X = \sum_{i=1}^{n} X_i$. Argue that each X_i is an independent random variable with a geometric distribution.
   Once we get $i-1$ coupons, there's a chance of $\frac{n-i+1}{n}$ chance of getting a coupon not already collected, if not we would keep on trying until we get a new coupon, thus X_i is a geometric random variable where
$$\text{Each } X_i \sim \text{Geom}\left(\frac{n-i+1}{n}\right)$$

b. Prove that $\mathbb{E}[X] \approx n \ln n$. Remember that the expectation of $\text{Geom}(p)$ is $\frac{1}{p}$.

We find the expectation of X using linearity of expectation and the expectation of a geometric random variable. Then we use the fact that $\sum_{i=1}^{n} \frac{1}{n} \approx \ln n$ (This sum is bounded both below and above by $\int_{x=1}^{n} \frac{dx}{x} + c$ where c is a constant).

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \sum_{i=1}^{n} \frac{n}{n-i+1} \\
&= n \sum_{i=1}^{n} \frac{1}{i} \\
&\approx n \ln n
\end{aligned}$$