HW4 Solutions

1. (20 pts.) Packets Over the Internet Again
n packets are sent over the Internet (n even). Consider the following probability models for the process:

(a) Each packet is routed over a different path and is lost independently with probability p.

(b) All n packets are routed along the same path and with probability p, one of the links along the path fails and all n packets are lost. Otherwise all packets are received.

(c) The n packets are divided into 2 groups of $n/2$ packets, and each group is routed along a different path and lost with probability p. Losses of different groups are independent events.

In each of the three models, compute the distribution, mean and variance of the number of packet losses. For $n = 6$ and $p = 0.3$, plot the distribution in each of the three cases. Does the distribution mean and variance depend on the probability model? Which of the three routing protocols do you prefer?

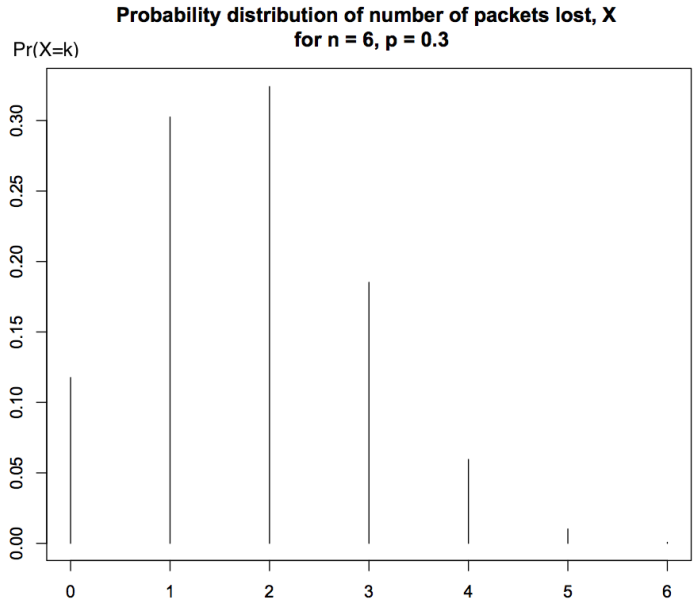
Answer: Let X be the random variable which denotes the number of packets lost.

(a) **(6 pts.)** Since each of the n packets is independently lost with probability p, we know that $X \sim \text{Bin}(n, p)$. Hence, for $k \in \{0, 1, \ldots, n\}$,

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

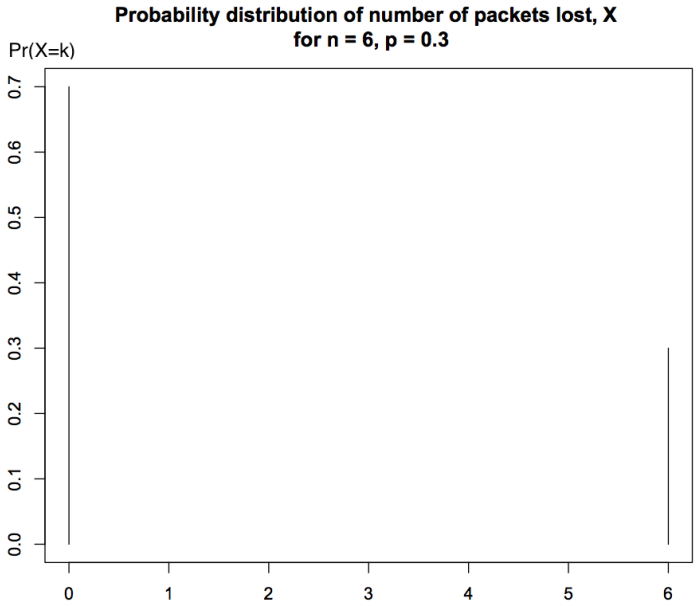
$$\mathbb{E}[X] = np$$

$$\text{Var}(X) = np(1-p)$$



(b) **(6 pts.)** In this case, either all or none of the packets are lost. So $\mathbf{P}(X = 0) = 1-p$ and $\mathbf{P}(X = n) = p$. Hence, $\mathbb{E}[X] = 0 \times (1-p) + n \times p = np$.

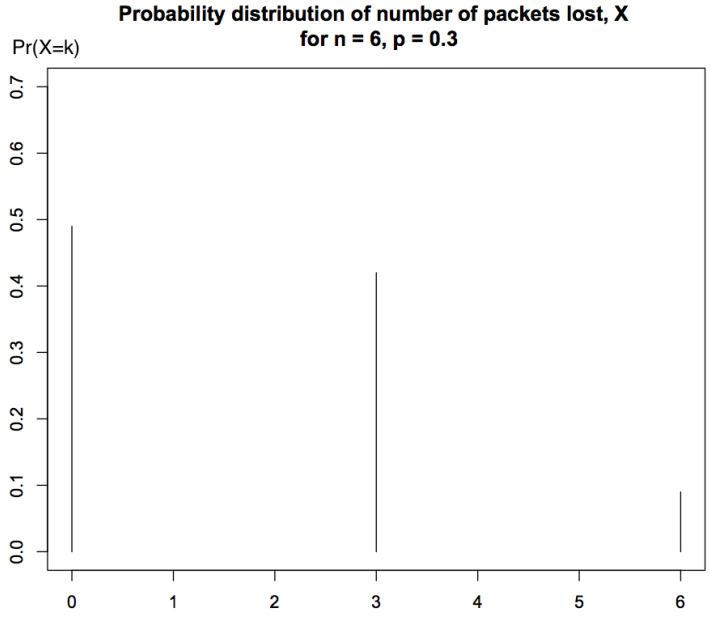
$$\mathrm{Var}(X) = n^2 p(1-p)$$



(c) **(6 pts.)** In this case, either no packets, exactly $n/2$ packets, or all n packets are lost. When $n/2$ packets are lost, it could be because path 1 failed while path 2 worked, or because path 2 failed while path 1 worked. Let Y_i represent whether path i fails ($Y_i = 1$ if it fails, $Y_i = 0$ if it does not). Then $X = (n/2)(Y_1 + Y_2)$ and $Y_i \sim \mathrm{Bern}(p)$. Hence, we get

$$\begin{aligned}
&\mathbf{P}(X = 0) = (1-p)^2 \\
&\mathbf{P}(X = n/2) = 2p(1-p) \\
&\mathbf{P}(X = n) = p^2 \\
&\mathbb{E}[X] = (n/2)\mathbb{E}[Y_1 + Y_2] = (n/2)(\mathbb{E}[Y_1] + \mathbb{E}[Y_2]) = np \\
&\mathrm{Var}(X) = (n/2)^2 \mathrm{Var}(Y_1 + Y_2) = (n/2)^2(\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)) = (n^2/2)p(1-p)
\end{aligned}$$

where the last line is due to the independence of Y_1, Y_2.



(2 pts.) As can be seen from the probability distribution tables and plots, the distributions and the variance depend on the probability models of packet loss, even though the expected number of lost packets is the same in all three models.

Which protocol is most preferable is quite dependent on the context, and it's possible to construct situations where each of the protocols performs best:

- If the packets were being transmitted without using an error-correcting code, then all of them are needed at the recipient's end to reconstruct the whole message. In this case, protocol (a) succeeds with probability $(1-p)^n$, protocol (b) with probability $1-p$ and protocol (c) with probability $(1-p)^2$. Therefore, protocol (b), which routes all packets on the same path, would work best in this situation.
- Suppose you used an error-correcting code that can tolerate the loss of up to 35% of packets, and suppose $n = 100$ and $p = 0.3$. Then protocol (a) succeeds with probability ≈ 0.884, protocol (b) with probability 0.7, and protocol (c) with probability 0.49. Therefore, protocol (a) works best in this situation.
- Suppose you used an error-correcting code that can tolerate the loss of up to 50% of packets, and suppose $n = 3$ and $p = 0.5$. Then protocol (a) and (b) succeed with probability 0.5, while protocol (c) succeeds with probability 0.75. In this case, protocol (c) works best.
- Finally, suppose that each packet that makes it to the recipient is worth the same amount; the total value of the communication is proportional to the number of packets received. Then it might not matter which protocol you choose: in the long run, they will deliver the same value, on average.

Any of these sorts of answers would be acceptable, as long as you give a coherent justification for which protocol you said is preferable.

Comment: There is a particularly important and common scenario where protocol (a) performs best by a wide margin. Suppose we know the packet loss probability p. Then if the number of packets n is large, it turns out that the actual number of packets lost will be pretty close to np (it will almost always be between $n(p-\epsilon)$ and $n(p+\epsilon)$, for some small constant ϵ). Hence, in such a case, it makes sense to use an error-correcting code which is designed to correct up to a $p+\epsilon$ of lost packets. By doing so, protocol (a) almost always succeeds in delivering enough packets that the error-correcting code can make up for the ones that were lost. This success probability can be made as close to 1 as desired (by increasing n). On the other

hand, protocol (b) will drop all the packets with probability p and protocol (c) with probability p^2, so they are considerably worse in this situation.

2. (20 pts.) Coupon Collection Again
Consider the coupon collection problem with n distinct baseball cards and m cereal boxes bought. Compute the distribution, mean and variance of the number of Babe Ruth cards acquired. (Hint: Think and write in terms of random variables.)

Answer: Each box of cereal is analogous to a biased coin flip. We either get Babe Ruth (coin comes up heads) w.p. $\frac{1}{n}$ or we don't w.p. $1 - \frac{1}{n}$ (coin comes up tails). Since each box of cereal is independent of all others then

$$X \sim Bin(m, 1/n)$$

where X is the random variable denoting the number of cards Babe Ruth cards acquired.

If we define the indicator random variable X_i such that $X_i = 1$ if we get Babe Ruth in box i and $X_i = 0$ otherwise then $X = \sum_{i=1}^{m} X_i$. Using the linearity of expectation we have

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{m} X_i\right] \\
&= \sum_{i=1}^{m} \mathbb{E}\left[X_i\right] \\
&= \sum_{i=1}^{m} \left[1 \cdot \mathbf{P}(X_i = 1) + 0 \cdot \mathbf{P}(X_i = 0)\right] \\
&= \sum_{i=1}^{m} \frac{1}{n} \\
&= \frac{m}{n}
\end{aligned}$$

By the independence of $X_1, \ldots, X_m$,

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{m} \mathrm{Var}(X_i) \\
&= \frac{m}{n}\left(1 - \frac{1}{n}\right)
\end{aligned}$$

It is a general fact that a random variable distributed $Bin(n, p)$ has expectation np and variance $np(1-p)$ (which is what we have shown here).

3. (20 pts.) DNA Sequencing Again
Compute expectation of the number of position on the genome sequence not covered by any reads

Answer: Let X be the number of positions not covered by any reads. Let X_i be an indicator random variable such that $X_i = 1$ if position i is not covered and $X_i = 0$ otherwise. Therefore, $X = \sum_{i=1}^{G} X_i$ so

using the linearity of expectation we have

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{G} X_i\right] \\
&= \sum_{i=1}^{G} \mathbb{E}[X_i] \\
&= \sum_{i=1}^{G} [\mathbf{P}(X_i = 1) \cdot 1 + \mathbf{P}(X_i = 0) \cdot 0] \\
&= \sum_{i=1}^{G} \left(1 - \frac{L}{G}\right)^N \\
&= G\left(1 - \frac{L}{G}\right)^N
\end{aligned}$$

4. (20 pts.) Family Planning

Mr. and Mrs. Brown decided to continue having children until they either have their first boy or until they have five children. Assume that each child is equally likely to be a boy or a girl, independent of all other children, and that there are no multiple births. Let B and G denote the numbers of boys and girls respectively that the Browns have.

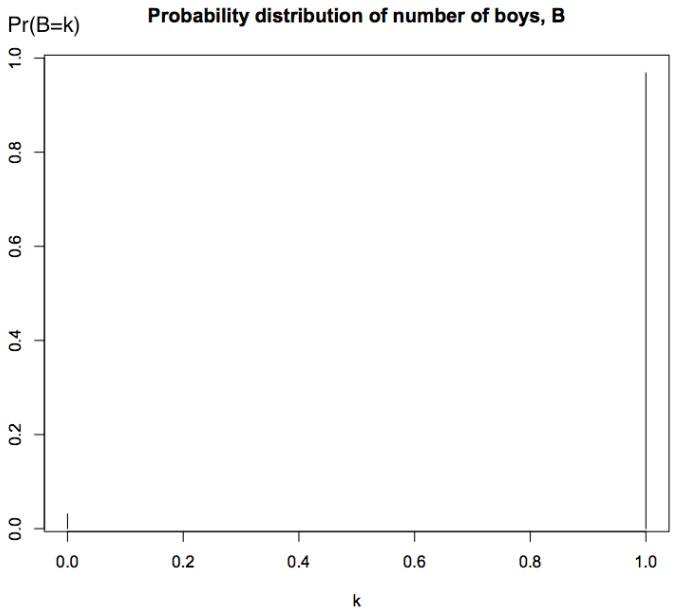
(a) **(4 pts.)** Write down the sample space together with the probability of each outcome

Answer: $\Omega = \{b, gb, ggb, gggb, ggggb, ggggg\}$, where b represents boy and g represents girl in the strings in Ω. For any $\omega \in \Omega$, let $\ell(\omega)$ represent the length of the string (i.e. the number of children born to the Browns). Then, $\mathbf{P}(\omega) = (\frac{1}{2})^{\ell(\omega)}$, since the gender of all children are independent and each gender is equally likely for every child.

(b) **(8 pts.)** Compute and plot the distributions of the random variables B and G.

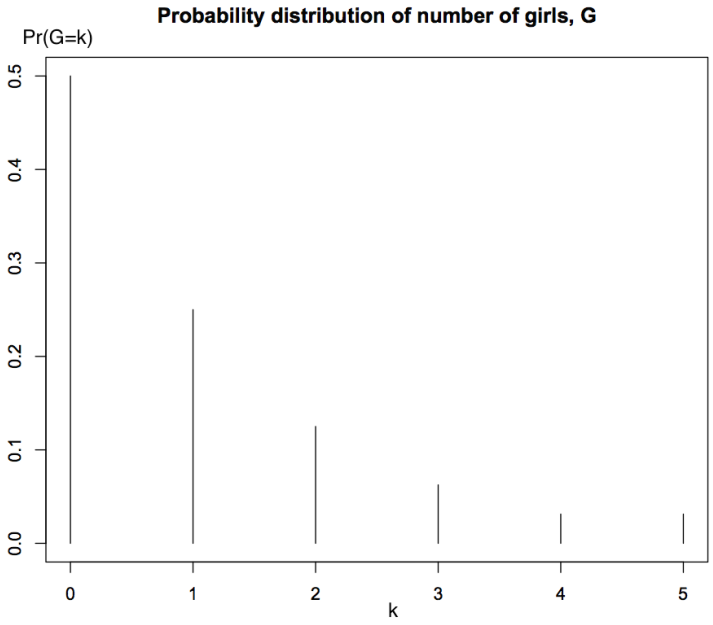
Answer:

$$\begin{aligned}
\mathbf{P}(B = 0) &= \frac{1}{2^5} \\
\mathbf{P}(B = 1) &= 1 - \frac{1}{2^5}
\end{aligned}$$



The distribution of G is as follows:

$$\mathbf{P}(G=0)=\frac{1}{2} \qquad \mathbf{P}(G=1)=\frac{1}{2^2}$$
$$\mathbf{P}(G=2)=\frac{1}{2^3} \qquad \mathbf{P}(G=3)=\frac{1}{2^4}$$
$$\mathbf{P}(G=4)=\frac{1}{2^5} \qquad \mathbf{P}(G=5)=\frac{1}{2^5}$$



(c) **(8 pts.)** Compute the mean and variance of B and G using a direct calculation

Answer:

$$\mathbb{E}[B]=\mathbf{P}(B=0)\cdot 0+\mathbf{P}(B=1)\cdot 1=\mathbf{P}(B=1)=\frac{31}{32}$$

$$\mathrm{Var}(B) = \mathbf{P}(B=0)\left(0-\frac{31}{32}\right)^2 + \mathbf{P}(B=1)\left(1-\frac{31}{32}\right)^2 = \left(\frac{1}{32}\right)\left(\frac{31}{32}\right)^2 + \left(\frac{31}{32}\right)\left(\frac{1}{32}\right)^2 = \frac{31}{1024}$$

$$\mathbb{E}[G] = \sum_{k=0}^{6} k\mathbf{P}(G=k) = \frac{1}{4} + \frac{2}{8} + \frac{3}{16} + \frac{4}{32} + \frac{5}{32} = \frac{31}{32}$$

$$\begin{aligned}\mathrm{Var}(G) = {} & \mathbf{P}(G=0)\left(0-\frac{31}{32}\right)^2 + \mathbf{P}(G=1)\left(1-\frac{31}{32}\right)^2 + \mathbf{P}(G=2)\left(2-\frac{31}{32}\right)^2 \\ & + \mathbf{P}(G=3)\left(3-\frac{31}{32}\right)^2 + \mathbf{P}(G=4)\left(4-\frac{31}{32}\right)^2 + \mathbf{P}(G=5)\left(5-\frac{31}{32}\right)^2 \\ = {} & \frac{1695}{1024}\end{aligned}$$

Comment: Note that $\mathbb{E}[B] = \mathbb{E}[G]$ above, which is a little counterintuitive. We might expect there to be more girls on average than boys, because the Browns might end up having many girls before their first boy. Or some might expect there to be more boys on average than girls, since the Browns will keep having children until they obtain a boy: thus they are guaranteed to have a boy, but there are no guarantees about girls. However, it turns out that the Browns' strategy actually does not change the expected number of boys vs girls. In fact, it doesn't matter when the family stops having children or what decision criteria they use; we will always have $\mathbb{E}[B] = \mathbb{E}[G]$, as long as each birth is equally likely to be a boy or girl.

5. (20 pts.) Function of a Random Variable

Let X be a random variable with the pmf $\mathbf{P}_X$ and $Y = g(X)$ be another random variable. Recall that $\mathbb{E}[Y]$ is defined to be $\sum_b b\mathbf{P}_Y(b)$, where $\mathbf{P}_Y$ is the pmf of Y. In this question we will verify the intuitive statement made in class:

$$\mathbb{E}[Y] = \sum_a g(a)\mathbf{P}_X(a), \tag{1}$$

(a) **(8 pts.)** First consider the example where X is uniformly distributed in $\{-n, -n+1, \ldots, n-1, n\}$

(i) **(4 pts.)**Show that eq. (1) holds if $Y = 2X$

Answer: From the problem statement we have

$$\mathbf{P}_X(a) = \frac{1}{2n+1} \quad a \in \{-n, \ldots, n\}$$

It is then apparent that the pmt of Y is given by

$$\mathbf{P}_Y(b) = \frac{1}{2n+1} \quad b \in \{-2n, -2n+2, \ldots, 2n-2, 2n\}$$

So from the definition of expectation

$$\begin{aligned}\mathbb{E}[Y] &= \frac{1}{2n+1}\left(-2n + (-2n+2) + \cdots + (2n-2) + 2n\right) \\ &= 0\end{aligned}$$

Using the statement from class,

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{a=-n}^{n} g(a)\mathbf{P}_X(a) \\ &= \sum_{a=-n}^{n} 2a\left(\frac{1}{2n+1}\right) \\ &= \frac{2}{2n+1}\sum_{a=-n}^{n} a \\ &= 0\end{aligned}$$

(ii) **(4 pts.)** Show that eq. (1) hold if $Y = X^2$

Answer: For ease of notation we we will denote $\mathcal{A} = \{0, 1, \ldots, (n-1)^2, n^2\}$ (this is referred to as the support of the PMF of Y). The PMF of Y can easily be found as

$$\mathbf{P}_Y(b) = \begin{cases} \frac{2}{2n+1} & b \in \mathcal{A}\backslash\{0\} \\ \frac{1}{2n+1} & b = 0 \\ 0 & \text{else} \end{cases}$$

Using the definition we have

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{b\in\mathcal{A}} bP_Y(b) \\ &= \frac{1}{2n+1}\cdot 0 + \frac{2}{2n+1}\left(1+4+9+\cdots+(n-1)^2+n^2\right)\end{aligned}$$

Faulhaber's formula says that the sum of squares of nonegative integers is $\frac{n(n+1)(2n+1)}{6}$. Plugging in gives

$$\mathbb{E}[Y] = \frac{n(n+1)}{3}$$

Now using the statement from class we can compute the expected value as follows:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{a=-n}^{n} g(a)P_X(a) \\ &= \frac{1}{2n+1}\sum_{a=-n}^{n} a^2 \\ &= \frac{1}{2n+1}\cdot 0 + \frac{2}{2n+1}\sum_{a=1}^{n} a^2 \\ &= \frac{2}{2n+1}\frac{n(n+1)(2n+1)}{6} \\ &= \frac{n(n+1)}{3}\end{aligned}$$

As an aside, $\mathbb{E}[X^2]$ is called the *second moment* of the distribution and is important for computing the variance of a random variable

(b) **(12 pts.)** Give a proof of eq. (1) in the general case

Answer:

$$\begin{aligned}
\mathbb{E}[Y] &\triangleq \sum_b b\mathbf{P}_Y(b) \\
&= \sum_b b\mathbf{P}(g(X) = b) \\
&= \sum_b b \sum_{\{a|g(a)=b\}} \mathbf{P}(X = a) \\
&= \sum_b \sum_{\{a|g(a)=b\}} b\mathbf{P}(X = a) \\
&= \sum_b \sum_{\{a|g(a)=b\}} g(a)\mathbf{P}(X = a) \\
&= \sum_a g(a)\mathbf{P}_X(a)
\end{aligned}$$