CS 70 Discrete Mathematics and Probability Theory
Summer 2019 James Hulett and Elizabeth Yang

HW 6

1 Trick or Treat

Ryan and Suraj are trick or treating together, and are each trying to collect all n flavors of Laffy Taffy. At each house, they each receive a Laffy Taffy, chosen uniformly at random from all flavors. However, Suraj will throw a tantrum if Ryan gets a flavor he doesn't, so they agree that if they receive different flavors, they'll both politely return the candy they got, and move on to the next house. What is the expected number of houses they need to visit until they get a full set of Laffy Taffy?

Solution:

Similar to the original coupon collector problem, we start by considering the expected number of houses that Ryan and Suraj need to visit to get a new flavor, once they already have i unique ones. At each house, there is a $\frac{n-i}{n}$ chance that Ryan gets a new flavor. However, in order for her to keep the candy, she needs Suraj to get the same flavor, which happens with probability $\frac{1}{n}$. Thus, the probability that Suraj and Ryan add a new flavor to their collection once they already have i distinct ones is $\frac{n-i}{n^2}$. This tells us that the expected number of houses they need to visit to get the $(i+1)$st flavor is $\frac{n^2}{n-i}$. Hence, in order to get all n flavors, Ryan and Suraj will have to visit on average

$$\begin{aligned}\sum_{i=0}^{n-1} \frac{n^2}{n-i} &= n^2 \cdot \sum_{i=0}^{n-1} \frac{1}{n-i} \\ &= n^2 \cdot \sum_{j=1}^{n} \frac{1}{j}\end{aligned}$$

where for the second equality, we did a variable substitution $j = n - i$ to simplify. As in the standard coupon collector problem, we know that the summation can be approximated by $\ln(n)$, so we have that in expectation, Ryan and Suraj have to visit approximately $n^2 \ln(n)$ houses until they have all n flavors.

2 Coupon Collection

Suppose you take a deck of n cards and repeatedly perform the following step: take the current top card and put it back in the deck at a uniformly random position (the probability that the card is placed in any of the n possible positions in the deck — including back on top — is $1/n$).

Consider the card that starts off on the bottom of the deck. What is the expected number of steps until this card rises to the top of the deck? (For large n, you may use the approximation $\sum_{i=1}^{n} \frac{1}{i} \approx \ln n$)

[*Hint*: Let T be the number of steps until the card rises to the top. We have $T = T_n + T_{n-1} + \cdots + T_2$, where the random variable T_i is the number of steps until the bottom card rises from position i to position $i-1$. Thus, for example, T_n is the number of steps until the bottom card rises off the bottom of the deck, and T_2 is the number of steps until the bottom card rises from second position to top position. What is the distribution of T_i?]

Solution:

Since a card at location i moves to location $i-1$ when the current top card is placed in any of the locations $i, i+1, \ldots, n$, it will rise with probability $p = (n-i+1)/n$. Thus, $T_i \sim \text{Geometric}(p)$, and $\mathbb{E}[T_i] = 1/p = n/(n-i+1)$. We now can see how this is exactly the coupon collector's problem, but with one fewer term (namely, without T_1). Finally, we can apply linearity of expectation to compute

$$\mathbb{E}[T] = \sum_{i=2}^{n} \mathbb{E}[T_i] = \sum_{i=2}^{n} \frac{n}{n-i+1} = n \sum_{i=2}^{n} \frac{1}{n-i+1} = n \sum_{i=1}^{n-1} \frac{1}{i} \approx n \ln(n-1).$$

3 Diversify Your Hand

You are dealt 5 cards from a standard 52 card deck. Let X be the number of distinct values in your hand. For instance, the hand (A, A, A, 2, 3) has 3 distinct values.

(a) Calculate $E[X]$.

(b) Calculate $Var[X]$.

Solution:

(a) Let X_i be the indicator of the ith value appearing in your hand. Then, $X = X_1 + X_2 + ... + X_{13}$ (Let 13 correspond to K, 12 correspond to Q, 11 correspond to J). By linearity of expectation then, $E[X] = \sum_{i=1}^{13} E[X_i]$. We can calculate $\mathbb{P}[X_i = 1]$ by taking the complement, $1 - Pr[X_i = 0]$, or 1 minus the probability that the card does not appear in your hand. This is $1 - \frac{\binom{48}{5}}{\binom{52}{5}}$. Then, $E[X] = 13\mathbb{P}[X_1 = 1] = 13(1 - \frac{\binom{48}{5}}{\binom{52}{5}})$.

(b) To calculate variance, since the indicators are not independent, we have to use the formula $E[X^2] = \sum_{i=j} E[X_i^2] + \sum_{i \neq j} E[X_i X_j]$.

$\sum_{i=j} E[X_i^2] = \sum_{i=j} E[X_i] = 13(1 - \frac{\binom{48}{5}}{\binom{52}{5}})$

To calculate $\mathbb{P}[X_i X_j = 1]$, we note that $\mathbb{P}[X_i X_j = 1] = 1 - \mathbb{P}[X_i = 0] - \mathbb{P}[X_j = 0] + \mathbb{P}[X_i = 0, X_j = 0]$.

$\sum_{i \neq j} E[X_i X_j] = 13 \cdot 12 \mathbb{P}[X_i X_j = 1] = 13 \cdot 12(1 - \mathbb{P}[X_i = 0] - \mathbb{P}[X_j = 0] + \mathbb{P}[X_i = 0, X_j = 0])$

$= 156(1 - 2\frac{\binom{48}{5}}{\binom{52}{5}} + \frac{\binom{44}{5}}{\binom{52}{5}})$

Putting it all together, we have $Var[X] = E[X^2] - E[X]^2 = 13(1 - \frac{\binom{48}{5}}{\binom{52}{5}}) + 156(1 - 2\frac{\binom{48}{5}}{\binom{52}{5}} + \frac{\binom{44}{5}}{\binom{52}{5}}) - (13(1 - \frac{\binom{48}{5}}{\binom{52}{5}}))^2$

4 Balls and Bins

Throw n balls into m bins, where m and n are positive integers. Let X be the number of bins with exactly one ball. Compute $\text{var}X$.

Solution:

Let X_i be the indicator that bin i has exactly one ball, for each $i = 1, \ldots, m$. Since $X = \sum_i X_i$, we can use the computational formula for variance:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{m} X_i\right)^2\right] - \left(\mathbb{E}[\sum_{i=1}^{m} X_i]\right)^2 \\
&= \mathbb{E}\left[\sum_{i \neq j} X_i X_j + \sum_{i=1}^{m} X_i^2\right] - \left(\sum_{i=1}^{m} \mathbb{E}[X_i]\right)^2 \\
&= \sum_{i \neq j} \mathbb{E}[X_i X_j] + \sum_{i=1}^{m} \mathbb{E}[X_i] - \left(\sum_{i=1}^{m} \mathbb{E}[X_i]\right)^2,
\end{aligned}$$

where the last line followed from linearity of expectation and recognizing that $X_i^2 = X_i$, since it can only take on the values 0 or 1.

One has

$$\mathbb{E}[X_i] = \binom{n}{1} \cdot \left(\frac{1}{m}\right)^1 \left(\frac{m-1}{m}\right)^{n-1} = \frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1}$$

and for $j \in \{1, \ldots, n\}$, $j \neq i$,

$$\mathbb{E}[X_i X_j] = \binom{n}{1}\binom{n-1}{1}\left(\frac{1}{m}\right)^1 \left(\frac{1}{m}\right)^1 \left(\frac{m-2}{m}\right)^{n-2} = \frac{n(n-1)}{m^2}\left(\frac{m-2}{m}\right)^{n-2}.$$

Noting that $\sum_{i \neq j}$ has $m(m-1)$ terms, and the rest of the sums have m terms, we find

$$\text{var}X = m(m-1) \cdot \frac{n(n-1)}{m^2}\left(\frac{m-2}{m}\right)^{n-2} + m \cdot \frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1} - m^2\left[\frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1}\right]^2.$$

5 Practical Confidence Intervals

(a) It's New Year's Eve, and you're re-evaluating your finances for the next year. Based on previous spending patterns, you know that you spend \$1500 per month on average, with a standard

deviation of $500, and each month's expenditure is independently and identically distributed. As a poor college student, you also don't have any income. How much should you have in your bank account if you don't want to go broke this year, with probability at least 95%?

(b) As a UC Berkeley CS student, you're always thinking about ways to become the next billionaire in Silicon Valley. After hours of brainstorming, you've finally cut your list of ideas down to 10, all of which you want to implement at the same time. A venture capitalist has agreed to back all 10 ideas, as long as your net return from implementing the ideas is positive with at least 95% probability.

Suppose that implementing an idea requires 50 thousand dollars, and your start-up then succeeds with probability p, generating 150 thousand dollars in revenue (for a net gain of 100 thousand dollars), or fails with probability $1-p$ (for a net loss of 50 thousand dollars). The success of each idea is independent of every other. What is the condition on p that you need to satisfy to secure the venture capitalist's funding?

(c) One of your start-ups uses error-correcting codes, which can recover the original message as long as at least 1000 packets are received (not erased). Each packet gets erased independently with probability 0.8. How many packets should you send such that you can recover the message with probability at least 99%?

Solution:

(a) Let T be the random variable representing the amount of money we spend in the year.

We have $T = \sum_{i=1}^{12} X_i$, where X_i represents the spending in the i-th month. So, $\mathbb{E}[T] = 12 \cdot \mathbb{E}[E_1] = 18000$.

And, since the X_is are independent, $\text{var}(T) = 12 \cdot \text{var}(X_1) = 12 \cdot 500^2 = 3,000,000$.

We want to have enough money in our bank account so that we don't finish the year in debt with 95% confidence. So, we want to keep some money ε more than the mean expenditure such that the probability of deviating above the mean by more than ε is less than 0.05.

Let's use Chebyshev's inequality here to express this.

$$\mathbb{P}(|T - \mathbb{E}[T]| \geq \varepsilon) \leq \frac{\text{var}(T)}{\varepsilon^2} \leq 0.05$$

This gives us $\varepsilon^2 \geq \frac{3,000,000}{0.05}$. So, $\varepsilon \geq 7746$. This means that we want to have a balance of $\geq \mathbb{E}[T] + \varepsilon = 25746$.

Observe that here, while we wanted to estimate $\mathbb{P}(T - \mathbb{E}[T] \geq \varepsilon)$, Chebyshev's inequality only gives us information about $\mathbb{P}(|T - \mathbb{E}[T]| \geq \varepsilon)$. But since

$$\mathbb{P}(|T - \mathbb{E}[T]| \geq \varepsilon) \geq \mathbb{P}(T - \mathbb{E}[T] \geq \varepsilon),$$

this is fine. We just get a more conservative estimate.

(b) For this question, to keep the numbers from exploding, let's work in thousands of dollars. Let X_i be the profit made from idea i, and T be the total profit made. We have $T = \sum_{i=1}^{10} X_i$.

Here, $\mathbb{E}[X_1] = 100p - 50(1-p) = 150p - 50$.

And $\text{var}(X_1) = 150^2 p(1-p)$ as the distribution of X_1 is a shifted and scaled Bernoulli distribution. Using $\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2$ yields the same answer.

We have, $\mathbb{E}[T] = 10 \cdot \mathbb{E}[X_1]$. Similarly, $\text{var}(T) = 10 \cdot \text{var}(X_1)$.

Now, we want to bound the probability of T going below 0 by 0.05. In other words, we want $\mathbb{P}(T < 0) \leq 0.05$.

But, in order to apply Chebyshev's inequality, we need to look at deviation from the mean. We use the assumption that to get our funding we obviously need $\mathbb{E}[T] > 0$. Then:

$$\mathbb{P}(T < 0) \leq \mathbb{P}(T \leq 0 \cup T \geq 2\,\mathbb{E}[T]) = \mathbb{P}(|T - \mathbb{E}[T]| \geq \mathbb{E}[T]) \leq \frac{\text{var}(T)}{\mathbb{E}[T]^2} \leq 0.05$$

Looking at just the last inequality, we have:

$$\frac{\text{var}(T)}{\mathbb{E}[T]^2} = \frac{10 \cdot \text{var}(X_1)}{100 \cdot \mathbb{E}[X_1]^2} = \frac{\text{var}(X_1)}{10 \cdot \mathbb{E}[X_1]^2} \leq 0.05$$

$$\therefore \frac{\text{var}(X_1)}{\mathbb{E}[X_1]^2} \leq 0.5$$

Now, substituting what we have for variance and expectation, we get the following:

$$-22500p^2 + 22500p \leq 0.5(150p - 50)^2$$

which gives us the quadratic:

$$33750p^2 - 30000p + 1250 \geq 0$$

The solutions for p are $p \geq \frac{1}{9}(4 + \sqrt{13})$ and $p \leq \frac{1}{9}(4 - \sqrt{13})$. So $p \geq 0.845$ or ≤ 0.0438.

The relevant solution here is to pick $p \geq 0.845$, since the other solution yields negative expectation (contradicting the earlier assumption of positive expectation).

(c) We want $k = 1000$ packets to get across without being erased. Say we send n packets. Let X_i be the indicator random variable representing whether the ith packet got across or not.

Let the total number of unerased packets sent across be T. We have $T = \sum_{i=1}^{n} X_i$ and we want $T \geq 1000$.

We want $\mathbb{P}(T < 1000) \leq 0.01$. Now, let's try to get this in a form so that we can use Chebyshev's inequality. We know that $\mathbb{E}[T] > 1000$, so we can say that

$$\begin{aligned}\mathbb{P}(T < 1000) &\leq \mathbb{P}\big(T \leq 1000 \cup T \geq \mathbb{E}[T] + (\mathbb{E}[T] - 1000)\big) \\ &= \mathbb{P}\big(|T - \mathbb{E}[T]| \geq (\mathbb{E}[T] - 1000)\big) \leq \frac{\text{var}(T)}{(\mathbb{E}[T] - 1000)^2} \leq 0.01.\end{aligned}$$

What is $\mathbb{E}[T]$? $\mathbb{E}[T] = n\,\mathbb{E}[X_1] = n(1-p) = 0.2n$.

Next, what is $\text{var}(T)$? $\text{var}(T) = n\,\text{var}(X_1) = np(1-p) = 0.16n$.

Now, $\dfrac{\text{var}(T)}{(\mathbb{E}[T]-k)^2} \leq 0.01 \implies 16n \leq (0.2n - 1000)^2$. This gives us the quadratic:

$$0.04n^2 - 416n + 1000000 \geq 0$$

Solving the last quadratic, we get $n \geq 6629$ or $n \leq 3774$. Since the second inequality doesn't make sense for our situation, our answer is $n \geq 6629$.

6 Law of Large Numbers

Recall that the *Law of Large Numbers* holds if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| > \varepsilon\right) = 0.$$

In class, we saw that the Law of Large Numbers holds for $S_n = X_1 + \cdots + X_n$, where the X_i's are i.i.d. random variables. This problem explores if the Law of Large Numbers holds under other circumstances.

Packets are sent from a source to a destination node over the Internet. Each packet is sent on a certain route, and the routes are disjoint. Each route has a failure probability of $p \in (0,1)$ and different routes fail independently. If a route fails, all packets sent along that route are lost. You can assume that the routing protocol has no knowledge of which route fails.

For each of the following routing protocols, determine whether the Law of Large Numbers holds when S_n is defined as the total number of received packets out of n packets sent. Answer **Yes** if the Law of Large Number holds, or **No** if not, and give a brief justification of your answer. (Whenever convenient, you can assume that n is even.)

(a) **Yes** or **No**: Each packet is sent on a completely different route.

(b) **Yes** or **No**: The packets are split into $n/2$ pairs of packets. Each pair is sent together on its own route (i.e., different pairs are sent on different routes).

(c) **Yes** or **No**: The packets are split into 2 groups of $n/2$ packets. All the packets in each group are sent on the same route, and the two groups are sent on different routes.

(d) **Yes** or **No**: All the packets are sent on one route.

Solution:

(a) **Yes.** Define X_i to be 1 if a packet is sent successfully on route i. Then X_i, $i = 1, \ldots, n$ is 0 with probability p and 1 otherwise. Since we have individual routes for each packet, we have a total of n routes. The total number of successful packets sent is hence $S_n = X_1 + \cdots + X_n$. Since S_n is a sum of i.i.d. Bernoulli random variables, $S_n \sim \text{Binomial}(n, 1-p)$.

Now similar to notation in the lecture notes, we define $A_n = S_n/n$ to be the fraction of successful packets sent, out of the n packets. Moreover, for each X_i,

$$\mathbb{E}[X_i] = 1 - p$$

and

$$\text{var}(X_i) = p(1-p).$$

Using Chebyshev's inequality:

$$\mathbb{P}[\,|A_n - \mathbb{E}[A_n]| > \varepsilon] = \mathbb{P}[\,|A_n - (1-p)| > \varepsilon] \leq \frac{\text{var}[A_n]}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

(b) **Yes.** Now we need $n/2$ routes for each pair of packets. Similarly to the previous question, we define X_i, $i = 1, \ldots, n/2$ to be 0 with probability p and 2 (packets) otherwise. Now the total number of packets is $S_n = X_1 + \cdots + X_{n/2}$ and the fraction of received packets is $A_n = S_n/n$.

Now for each $i = 1, \ldots, n/2$,

$$\mathbb{E}[X_i] = 2(1-p)$$

and

$$\text{var}(X_i) = 4p(1-p).$$

Thus,

$$\mathbb{E}[A_n] = \frac{\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_{n/2}]}{n} = \frac{1}{n} \cdot \frac{n}{2} \cdot 2(1-p) = 1 - p$$

and

$$\text{var}[A_n] = \frac{1}{n^2}\left(\text{var}[X_1] + \ldots + \text{var}[X_{n/2}]\right) = \frac{1}{n^2} \cdot \frac{n}{2} 4p(1-p) = \frac{2p(1-p)}{n}.$$

Finally, we get:

$$\mathbb{P}[\,|A_n - \mathbb{E}[A_n]| > \varepsilon] = \mathbb{P}[\,|A_n - (1-p)| > \varepsilon] \leq \frac{2p(1-p)}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

(c) **No.** In this situation, we have that no packets get through with probability p^2, half the packets get through with probability $2p(1-p)$, and all the packets get through with probability $(1-$

$p)^2$. This tells us that $\frac{1}{n}S_n$ is 0 with probability p^2, $\frac{1}{2}$ with probability $2p(1-p)$, and 1 with probability $(1-p)^2$. Since $\mathbb{E}\left[\frac{1}{n}S_n\right] = 1-p$, this gives us that

$$\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| = \begin{cases} 1-p & \text{with probability } p^2 \\ \left|p-\frac{1}{2}\right| & \text{with probability } 2p(1-p) \\ p & \text{with probability } (1-p)^2 \end{cases}$$

We now consider two cases: either $p = \frac{1}{2}$ or $p \neq \frac{1}{2}$. In the former case, we can take $\varepsilon = \frac{1}{4}$, and we'll have that

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| > \varepsilon\right) &= \mathbb{P}\left(\frac{1}{n}S_n = 0 \cup \frac{1}{n}S_n = 1\right) \\ &= \frac{1}{2} \end{aligned}$$

In the latter case, we can take $\varepsilon = \frac{\min\left(1-p, \left|p-\frac{1}{2}\right|, p\right)}{2}$ and we'll have that

$$\mathbb{P}\left(\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| > \varepsilon\right) = 1$$

Since neither of these probabilities converge to zero as $n \to \infty$, we have that the WLLN does not hold in either case.

(d) **No.** In this case, we have that no packets get through with probability p and all the packets get through with probability $(1-p)$. Hence,

$$\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| = \begin{cases} 1-p & \text{with probability } p \\ p & \text{with probability } (1-p) \end{cases}$$

So if we take $\varepsilon = \frac{\min(p,1-p)}{2}$, we have that

$$\mathbb{P}\left(\left|\frac{1}{n}S_n - \mathbb{E}\left[\frac{1}{n}S_n\right]\right| > \varepsilon\right) = 1$$

As in the previous part, because this does not converge to 0 as $n \to \infty$, we have that the WLLN does not hold.

For problems (c) and (d), you should've had the intuition that since the packets are automatically sent through 1 or 2 routes, increasing n does not really help for LLN.