Homework 12

This homework is due November 24, 2014, at 12:00 noon.

1. **Section Rollcall!**

   In your self-grading for this question, give yourself a 10, and write down what you wrote for parts (a) and (b) below as a comment. Please put the answers in your written homework as well.

   (a) What discussion did you attend on Monday last week? If you did not attend section on that day, please tell us why.

   (b) What discussion did you attend on Wednesday last week? If you did not attend section on that day, please tell us why.

2. **Hashing & Drunk Man Lab**

   (Note): This Virtual Lab will be pushed to Homework 13. If you want to get started early on it, that's great! If not, just hand it in with Homework 13.

   The Birthday Paradox, which we explored in last week's lab, is a specific form of a more general concept known as hashing. In hashing, we are interested in minimizing the probability that two or more keys are hashed to the same table location. This also can be looked at as a balls and bins problem, where we are trying to minimize the probability of balls being thrown into the same bin.

   Let A be the event that there is no collision (i.e. no two balls are thrown into the same bin), and $P(A)$ be the probability that there is no collision. For the questions below, let's consider n bins and m balls (for the Birthday Paradox, $n = 365$).

   For each part, students who want to can choose to completely rewrite the question. Basically, you can come up with your own formulation of how to do a series of experiments that result in the same discoveries. Then, write up the results nicely using plots as appropriate to show what you observed. You can also rewrite the entire lab to take a different path through as long as they convey the key insights aimed at in each part.

   Please download the IPython starter code from Piazza or the course webpage, and answer the following questions.

   (a) For $n = 100, 365, 500, 1000$, and for $m = 0.1n, 0.3n, \sqrt{n}$, simulate throwing m balls into n bins. For every (m, n) pair, do this for 1000 trials. For each value of n, plot the fraction of trials for which there is no collision vs. m. What do you observe as m gets larger?

   (b) Let's take a detour and look at a very useful approximation for $ln(1-x)$. Plot $ln(1-x)$ for 1000 values of x from $0.01 - 0.1$. What is the approximate slope of this graph and why? (think about Taylor expansion.)

   (c) Repeat simulating throwing m balls in n bins for $n = 100, 365, 500, 1000$, and 20 values of m between 1 and $2\sqrt{n}$. Now for each n, plot the negative logarithm of the fraction of trials for which there were no collisions, for different values of m. Does this plot look linear in m? Quadratic? Exponential?

(d) Now overlay the plots in part (c) with $\frac{m^2}{2n}$. What do you notice?

From these graphs, and for each value of n, find the approximate m value such that $P(A) \approx 0.5$ for both the simulated data and $\frac{m^2}{2n}$. How much does this differ?

(e) The drunk man is back, and he wants your help to plot 100 sample paths that he could take from time $t = 0$ to $t = 999$ (1000 timesteps). Assume the same probabilities from Homework 11's Question 5, that is, the man moves forward with probability 0.5, backward with probability 0.3, and stays exactly where he is with probability 0.2. What do you observe about his paths? Where do you think he should end up at after 1000 timesteps, on average?

Hint: Implement the function `drunk_man`, which returns a list of elements that starts at 0, and every element thereafter is one more, one less, or equal to the previous one, with the correct probability for each possibility.

(f) Redo the previous part, but this time, the man moves forward with probability 0.3, backward with probability 0.5, and stays exactly where he is with probability 0.2. What do you observe about his paths? Where do you think he should end up at after 1000 timesteps, on average?

Finally, assume the man moves forward and backward with probability 0.5 each, i.e. he never stays still. Plot his 100 sample paths. What do you observe? Does this graph remind you of something we have done in a previous virtual lab?

(g) One simple way of solving counting problems is to enumerate all possibilities, and then count the ones that we are looking for. In this question, let's check your answer to Question 5 from the last homework. Implement the function `count_paths(t)`, which generates all the paths the Drunk Man can take in t timesteps, then counts the number of paths in which he returns to 0 at time t and it is his first return. Remember that we no longer care about probabilities when counting paths.

Hint: One way to generate all possible paths is to use `itertools.product`. You might also want to implement the function `catalan(n)`, which computes the n^{th} Catalan number.

(h) The Bernoulli distribution is the probability distribution of a random variable which takes value 1 with success probability p and value 0 with failure probability $1 - p$. It can be used, for example, to represent a coin toss, where 1 is defined to mean "heads" and 0 is defined to mean "tails".

In other words:

$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

Related values:

$$E[X] = p$$
$$Var(X) = p(1-p)$$

Plot the probability mass functions (pmf) for Bernoulli random variables with success probabilities of $0.1, 0.3, 0.5, 0.7$, and 0.9, respectively.

Hint: `scipy.stats.bernoulli.pmf` will be useful for this question.

(i) The binomial distribution gives us the probability of observing k successes, each with a probability p, out of N attempts.

$$f(k;N;p) = \binom{N}{k} p^k (1-p)^{N-k}$$

where

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

with $k = 0, 1, 2, \ldots, N$
Related values:

$$E[X] = Np$$

$$Var(X) = Np(1-p)$$

If I toss a 0.6-biased coin 20 times, what is the approximate probability of getting exactly 15 heads? Answer this question graphically by plotting the PMF of a binomial random variable with parameters $(20, 0.6)$.
Looking at the plot, at which value of n is the probability of getting n heads maximized?
Hint: As before, you will find `scipy.stats.binom.pmf` useful.

(j) Question 9, part (e)

(k) Question 9, part (f)

Reminder: When you finish, don't forget to convert the notebook to pdf and merge it with your written homework. Please also zip the `ipynb` file and submit it as `hw12.zip`.

3. Fundamentals
True or false? For the following statements, provide either a proof or a simple counterexample. Let X, Y, Z be arbitrary random variables.

(a) If (X,Y) are independent and (Y,Z) are independent, then (X,Z) are independent.
FALSE. Let X,Y be i.i.d Bernoulli$(1/2)$ random variables, and let $Z = X$. Then (X,Y) and (Y,Z) are independent by construction, but (X,Z) are not independent because $Z = X$.

(b) If (X,Y) are dependent and (Y,Z) are dependent, then (X,Z) are dependent.
FALSE. Let X,Z be i.i.d Bernoulli$(1/2)$ random variables, and let $Y = XZ$. Then (X,Z) are independent by construction, but (X,Y) are not independent, since $\Pr[X = 0 \wedge Y = 1] = 0 \neq \Pr[X = 0]\Pr[Y = 1] = (1/2)(1/4)$.

(c) Assume X is discrete. If $\mathrm{Var}(X) = 0$, then X is a constant.
TRUE. Let $\mu = \mathbf{E}[X]$. By definition,

$$0 = \mathrm{Var}(X) = \mathbf{E}[(X-\mu)^2] = \sum_{\omega \in \Omega} \Pr[\omega](X(\omega) - \mu)^2$$

The RHS is the sum of non-negative numbers, so if the sum is 0, each term must be 0.
So $\Pr[\omega] > 0 \implies (X(\omega) - \mu)^2 = 0 \implies X(\omega) = \mu$. Therefore X is constant (equal to $\mu = \mathbf{E}[X]$).

(d) $\mathbf{E}[X]^4 \leq \mathbf{E}[X^4]$
TRUE. First, for an arbitrary random variable Y, we have:

$$0 \leq \mathbf{E}[(Y - \mathbf{E}[Y])^2] = \mathbf{E}[Y^2] - \mathbf{E}[Y]^2$$

So $\mathbf{E}[Y]^2 \leq \mathbf{E}[Y^2]$. Now applying this twice, once for $Y = X$ and once for $Y = X^2$:

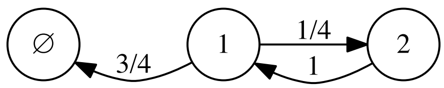
$$\mathbf{E}[X]^4 = (\mathbf{E}[X]^2)^2 \leq (\mathbf{E}[X^2])^2 \leq \mathbf{E}[(X^2)^2] = \mathbf{E}[X^4]$$

4. Markov Conversation

Alice is hosting a party. As she's talking to her guests, she notices that conversations naturally transition between casual and more interesting topics. Consider the following simple model of conversations: Each type of topic takes a certain amount of time, and can transition to different topics as specified.

1. Casual Topics: These take 5 minutes. At the end, they can transition into an Interesting topic (w.p. 25%), or the conversation can terminate (w.p. 75%).
2. Interesting Topics: These take 16 minutes. At the end, they are always followed by a Casual Topic.

The following diagram illustrates the conversation flow, where "∅" means the conversation has terminated, and "1", "2" correspond to the conversation topics.



Using the above model:

(a) What is the expected length of a conversation that starts on a Casual topic?

(Hint: Let the random variable X be the length of a conversation that starts on a Casual topic. Try to express $\mathbf{E}[X]$ *in terms of conditional expectations...)*

Let L be the event that after the (initial) Casual topic, the conversation terminates. So $P(L) = 3/4$. Let R be the event that after the (initial) Casual topic, the conversation transitions into an Interesting topic. So $P(R) = 1/4$.

We clearly have $\mathbf{E}[X|L] = 5$, since given the conversation terminates after the initial topic, the expected length is just 5 minutes.

Now to compute $\mathbf{E}[X|R]$: Given the conversation transitions into an Interesting topic, it lasts 5 minutes (for the initial topic), plus 16 minutes (for the Interesting topic), plus some X' minutes once we're back at Casual topics (where X' is another random variable). What is the distribution of X'? Notice, X' is just the length of a conversation that starts on a Casual topic. Therefore X' is distributed identically as X! So $\mathbf{E}[X'] = \mathbf{E}[X]$, and we can write:

$$\mathbf{E}[X|R] = \mathbf{E}[5+16+X'|R] = 5+16+\mathbf{E}[X'|R] = 21+\mathbf{E}[X]$$

(Notice, the r.v. X' is independent of event R, since event R only concerns the **first** transition from the initial topic. Therefore $\mathbf{E}[X'|R] = \mathbf{E}[X'] = \mathbf{E}[X]$).

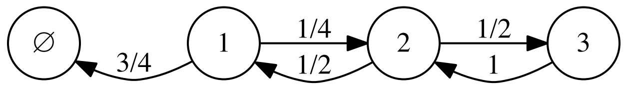
Now we can write $\mathbf{E}[X]$ using the conditional expectations:

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}[X|L]P(L) + \mathbf{E}[X|R]P(R) \\ &= 5(3/4) + (21+\mathbf{E}[X])(1/4) \\ &= 5 + 1/4(16+\mathbf{E}[X]) \\ \mathbf{E}[X] &= 9 + \frac{1}{4}\mathbf{E}[X]\end{aligned}$$

Solving for $\mathbf{E}[X]$ in the above equation, we find $\mathbf{E}[X] = 12$.

As the party progresses, Alice revises her model of conversation to include three types of topics:

1. Casual Topics: These take 5 minutes. At the end, they can transition into an Interesting topic (w.p. 25%), or the conversation can terminate (w.p. 75%).
2. Interesting Topics: These take 15 minutes. At the end, they can be followed by a Deep topic (w.p. 50%), or can go back to a Casual topic (w.p. 50%).
3. Deep Topics: These take 25 minutes. They are always followed by an Interesting topic.



Alice starts to wonder how the expected length of her conversations depend on who she talks to. Assume the following model: If she talks to acquaintances, they start on a Casual topic. With close friends, they start on an Interesting topic (w.p. 50%), or on a Deep topic (w.p. 50%). Using the revised model:

(b) Alice starts talking to her acquaintance Bob. What is the expected length of their conversation?

Hint: Let X_1, X_2, X_3 be the length of her conversation if she starts on topic $1, 2, 3$ respectively. Try to relate the expected values of the three R.Vs.

Considering conversations that start on topic 1, let L be the event that the conversation terminates after the initial topic, and let R be the event that it transitions to topic 2. We clearly have $\mathbf{E}[X_1|L] = 5$. Further, given it transitions to topic 2, the conversation takes 5 minutes, plus the length of a conversation that starts in topic 2. That is: $\mathbf{E}[X_1|R] = \mathbf{E}[5 + X_2|R] = 5 + \mathbf{E}[X_2]$, where $\mathbf{E}[X_2]$ is the expected length of a conversation that starts in topic 2. Putting this together (as in the previous part):

$$\begin{aligned}\mathbf{E}[X_1] &= \mathbf{E}[X_1|L]P(L) + \mathbf{E}[X_1|R]P(R)\\ &= 5(3/4) + (5 + \mathbf{E}[X_2])(1/4)\\ &= 5 + \frac{1}{4}\mathbf{E}[X_2].\end{aligned}$$

Finding similar relations for conversations that start on topics 2 and 3, we have a system of linear equations in $\mathbf{E}[X_1], \mathbf{E}[X_2], \mathbf{E}[X_3]$:

$$\begin{aligned}\mathbf{E}[X_1] &= 5 + \frac{1}{4}\mathbf{E}[X_2]\\ \mathbf{E}[X_2] &= 15 + \frac{1}{2}\mathbf{E}[X_1] + \frac{1}{2}\mathbf{E}[X_3]\\ \mathbf{E}[X_3] &= 25 + \mathbf{E}[X_2]\end{aligned}$$

We can solve this by simple substitution to find:
$\mathbf{E}[X_1] = 25$ minutes.
$\mathbf{E}[X_2] = 55 + \mathbf{E}[X_1] = 80$ minutes.
$\mathbf{E}[X_3] = 25 + \mathbf{E}[X_2] = 105$ minutes.
Since acquaintances always start on topic 1, the expected length of a conversation between Alice and Bob is $\mathbf{E}[X_1] = 25$ minutes.

(c) Alice starts talking to her close friend Charlie. What is the expected length of their conversation?

They start on topic 2 with probability $1/2$, and on topic 3 with probability $1/2$, so the expected length is $1/2\mathbf{E}[X_2]+1/2\mathbf{E}[X_3]=92.5$ minutes.

(d) Assume people at the party are equally likely to be close friends or acquaintances. But Eve noticed that Alice and Dave talked for 45 minutes! What is the probability that Dave is a close friend of Alice?

There are only two possible conversation paths that take 45 minutes:

- Path A: $1 \to 2 \to 1 \to 2 \to 1 \to \emptyset$
- Path B: $3 \to 2 \to 1 \to \emptyset$

So the event that the conversation lasts 45 minutes is exactly the event that the conversation took one of these two paths. Further, Path A starts on a Casual topic (which only acquaintances would start on), and Path B starts on a Deep topic (which only close friends would start on). Therefore, letting the event "Path A" be the event that the conversation took Path A, the probability we are interested in is:

$$\Pr[\text{Path B} \mid \text{Path A or Path B}]$$

It is easy to find Pr[Path A] if we first condition on the conversation starting on topic 1: Pr[Path A] = Pr[Path A $\cap$ start at 1] = Pr[Path A | start at 1]Pr[start at 1].

From the transition probabilities, we can find p_A = Pr[Path A | start at 1] = $(1/4)^2(1/2)^2(3/4)$, since each $1 \to 2$ transition occurs with probability $1/4$, each $2 \to 1$ transition occurs w.p. $1/2$, and $1 \to \emptyset$ occurs w.p. $3/4$.

And similarly, p_B = Pr[Path B | start at 3] = $(1/2)(3/4)$.

So we have:

$$\begin{aligned}\Pr[\text{Path B} \mid \text{Path A or Path B}] &= \frac{\Pr[\text{Path B} \mid \text{start at 3}]\Pr[\text{start at 3}]}{\Pr[\text{Path A} \mid \text{start at 1}]\Pr[\text{start at 1}] + \Pr[\text{Path B} \mid \text{start at 3}]\Pr[\text{start at 3}]} \\ &= \frac{p_B(1/4)}{p_A(1/2)+p_B(1/4)} = \frac{p_B}{2p_A+p_B} = 16/17 \approx 0.94\end{aligned}$$

5. Round the Clock

As your new year's resolution you decided to have three meals a day, with no exceptions. However you do not like putting names on things, and so you won't be calling your meals breakfast, lunch, and/or dinner. You simply love them equally and hate labeling them. Furthermore, towards your goal of being well-organized you decide that you will have each meal at the same time everyday.

For each one of your meals, you decide that you would have it at an exact but completely random hour during the whole day. i.e. for each meal you pick a time from the set of 24 exact hours $\{12\text{am}, 1\text{am}, \ldots, 10\text{pm}, 11\text{pm}\}$ and then you have that meal at that exact time every day. Note that you might even have two meals at the same time (i.e. you sample with replacement)! But if you schedule two or more meals at the same time, you decide on a completely random ordering over the meals scheduled for that time, and then you'll have your meals according to that ordering every day. Furthermore, for simplicity assume that it takes no time for you to finish a meal.

(a) For each one of your meals, what is the expected amount of time you have to wait after finishing it, until you get to have your next meal?

Let X_1, X_2, X_3 be the random variables, where X_i shows how long we have to wait until the next meal after we finish meal i. By linearity of expectation we have $\mathbf{E}(X_1+X_2+X_3) = \mathbf{E}(X_1)+\mathbf{E}(X_2)+\mathbf{E}(X_3)$.

But note that $X_1+X_2+X_3$ is always 24, because if at every point in the day there is exactly one meal that was finished last, and so that point in time could be attributed to that meal. This way we can see that entire 24 hours of a day are perfectly partitioned by the intervals we are waiting after meals 1, 2, and 3. So $X_1+X_2+X_3=24$. But then everything in the problem statement is symmetric about the meals. We don't even label them! So it must be true that X_1, X_2, and X_3 have the same expected value. So we get

$$24=\mathbf{E}(X_1+X_2+X_3)=\mathbf{E}(X_1)+\mathbf{E}(X_2)+\mathbf{E}(X_3)=3\times\mathbf{E}(X_1).$$

This means that $\mathbf{E}(X_1)=8$, and similarly $\mathbf{E}(X_2)=\mathbf{E}(X_3)=8$. So for each meal, we have to wait an expected 8 hours after finishing it for the next meal.

(b) Now suppose that your friend calls you at an exact random hour during the day (24 different possibilities). S/he asks you to go to his/her place and have your next meal with him/her. You do not want to ruin your schedule, so you tell your friend that you'll be at his/her place in X hours to have your next meal, where $X \geq 0$ is the number of hours until your next meal is scheduled. **What is the expected value of X?** (If a phone call falls on the exact hour of a meal, break ties randomly as in part (a)).

We use the following observation: the phone call is not that different from a meal! Its time is random, just like the meals, and even when its scheduled time is tied with some of the meals, the ordering of the meals and the phone call becomes perfectly uniformly random. Therefore if we think of a phone call as a fourth meal, then by following a similar reasoning as in the previous part we get that the expected time we have to wait after each meal until the next one is $\frac{24}{4}=6$. In particular, after having the fourth meal, the phone call, the expected amount of time we need to wait for the next meal is 6 hours.

Side note: Naively one might think that the answer should be half the answer to the previous part, because the in-between-meals interval in which the phone call happens has expected length 8 and so it might seem OK to say that on average the phone call happens in the middle of the interval, so the time from the phone call until the next meal is 4. But this reasoning is wrong, because when everything is random, there are in-between-meals intervals that are larger than 8 hours and there are also intervals smaller than 8 hours. If the phone call happens in intervals larger than 8 hours, then the average value for X would be larger than 4. One might think that the times the phone call happens in large intervals cancel out with the times the phone call happens in smaller intervals, but this is not true. A phone call is placed at a random time, and therefore it is more likely to fall into a larger interval, than in a small interval.

6. Linearity

Solve each of the following problems using linearity of expectation. Explain your methods clearly.

(a) In an arcade, you play game A 10 times and game B 20 times. Each time you play game A, you win with probability $1/3$ (independently of the other times), and if you win you get 3 tickets (redeemable for prizes), and if you lose you get 0 tickets. Game B is similar, but you win with probability $1/5$, and if you win you get 4 tickets. What is the expected total number of tickets you receive?

Let A_i be the indicator you win the i^{th} time you play game A and B_i be the same for game B. The expected value of A_i and B_i are,

$$\mathbf{E}[A_i]=1\cdot 1/3+0\cdot 2/3=1/3,$$
$$\mathbf{E}[B_i]=1\cdot 1/5+0\cdot 4/5=1/5.$$

Let T_A be the random variable for the number of tickets you win in game A, and T_B be the number of tickets you win in game B.

$$\mathbf{E}[T_A + T_B] = 3\mathbf{E}[A_1] + \cdots + 3\mathbf{E}[A_{10}] + 4\mathbf{E}[B_1] + \cdots + 4\mathbf{E}[B_{20}]$$
$$= 10\left(3 \cdot \frac{1}{3}\right) + 20\left(4 \cdot \frac{1}{5}\right) = 26$$

□

(b) A monkey types at a 26-letter keyboard with one key corresponding to each of the lower-case English letters. Each keystroke is chosen independently and uniformly at random from the 26 possibilities. If the monkey types 1 million letters, what is the expected number of times the sequence "book" appears?

There are $1,000,000 - 4 + 1 = 999,997$ places where "book" can appear, each with a (non-independent) probability of $\frac{1}{26^4}$ of happening. If A is the random variable that tells how many times "book" appears, and A_i is the indicator variable that is 1 if "book" appears starting at the i^{th} letter, then

$$\mathbf{E}[A] = \mathbf{E}[A_1 + \cdots + A_{999,997}]$$
$$= \mathbf{E}[A_1] + \cdots + E[A_{999,997}]$$
$$= \frac{999,997}{26^4} \approx 2.19$$

times. □

(c) A building has n floors numbered $1, 2, \ldots, n$, plus a ground floor G. At the ground floor, m people get on the elevator together, and each gets off at a uniformly random one of the n floors (independently of everybody else). What is the expected number of floors the elevator stops at (not counting the ground floor)?

Let A_i be the indicator that the elevator stopped at floor i.

$$\Pr[A_i = 1] = 1 - \Pr[\text{no one gets off at } i] = 1 - \left(\frac{n-1}{n}\right)^m.$$

If A is the number of floors the elevator stops at, then

$$\mathbf{E}[A] = \mathbf{E}[A_1 + \cdots + A_n]$$
$$= \mathbf{E}[A_1] + \cdots + E[A_n] = n \cdot \left(1 - \left(\frac{n-1}{n}\right)^m\right)$$

□

(d) A coin with Heads probability p is flipped n times. A "run" is a maximal sequence of consecutive flips that are all the same. (Thus, for example, the sequence HTHHHTTH with $n = 8$ has five runs.) Show that the expected number of runs is $1 + 2(n-1)p(1-p)$. Justify your calculation carefully.

Let A_i be the indicator for the event that a run starts at the i toss. Let $A = A_1 + \cdots + A_n$ be the random variable for the number of runs total. Obviously, $E[A_1] = 1$. For $i \neq 1$,

$$\mathbf{E}[A_i] = \Pr[A_i = 1]$$
$$= \Pr[i = H \mid i-1 = T] \cdot \Pr[i-1 = T] + \Pr[i = T \mid i-1 = H] \cdot \Pr[i-1 = H]$$
$$= p \cdot (1-p) + (1-p) \cdot p$$
$$= 2p \cdot (1-p).$$

This gives

$$\begin{aligned}\mathbf{E}[A] &= \mathbf{E}[A_1 + A_2 + \cdots + A_n] \\ &= \mathbf{E}[A_1] + \mathbf{E}[A_2] + \cdots + \mathbf{E}[A_n] = 1 + 2(n-1)p(1-p).\end{aligned}$$

□

7. Company Selection

Company A produces a particular device consisting of 10 components. Company A can either buy all the components from Company S or Company T, and then uses them to produce the devices without testing every individual component. After that, each device will be tested before leaving the factory. The device works only if every component works properly. Each working device can be sold for x dollars, but each non-working device must be thrown away. Products from Company S have a failure probability of $q = 0.01$ while Company T has a failure probability of $q/2$. However, every component from Company S costs \$10 while it costs \$30 from Company T. Should Company A build the device with components from Company S or Company T in order to maximize its expected profit per device? (Your answer will depend on x.)

Let W denote the event that a device works. Let R be the random variable denoting the profit.

$$\mathbf{E}[R] = P(W)\mathbf{E}[R|W] + P(W^C)\mathbf{E}[R|W^C].$$

Let's first consider the case when we use products from Company S. In this case, a device works with probability $P[W] = (1-q)^{10}$. The profit made on a working device is $x - 100$ dollars while a nonworking device has a profit of -100 dollars. That is, $\mathbf{E}[R|W] = x - 100$ and $\mathbf{E}[R|W^C] = -100$. Using R_S to denote the profit using components from Company S, the expected profit is:

$$\mathbf{E}[R_S] = (1-q)^{10}(x-100) + (1-(1-q)^{10})(-100) = (1-q)^{10}x - 100 = (0.99)^{10}x - 100.$$

If we use products from Company T. The device works with probability $P[W] = (1-q/2)^{10}$. The profit per working device is $\mathbf{E}[R|W] = x - 300$ dollars while the profit for a nonworking device is $\mathbf{E}[R|W^C] = -300$ dollars. The expected profit is:

$$\mathbf{E}[R_T] = (1-q/2)^{10}(x-300) + (1-(1-q/2)^{10})(-300) = (1-q/2)^{10}x - 300 = (0.995)^{10}x - 300.$$

To determine which Company should we use, we solve $\mathbf{E}[R_T] \geq \mathbf{E}[R_S]$, yielding $x \geq 200/[(0.995)^{10} - (0.99)^{10}] = 4280.1$. So for $x < \$4280.1$ using products from Company S results in greater profit, while for $x > \$4280.1$ more profit will be generated by using products from Company T.

8. (*Extra Credit*) Would You Bet It?

In addition to the oral exam, suppose EECS70 offers an extra credit challenge as follows.

- Bet: 4 homework points.
- Rules:
  - Initially, there are 52 cards with 4 suits (♠, ♡, ♢, ♣) and 13 values (A,2,3,4,5,6,7,8,9,10,J,Q,K).
  - In each round, you draw a card and put it in front of you. If two cards in front of you have the same value, the game is over; otherwise, you survive this round and earn 1 homework point.
  - All cards you drew stay in front of you until the game is over.

Would it be wise to play the game? You might want to use a spreadsheet program to help your calculation.

It is a good choice because the expected point from playing this game is $-4+4.6966 = 0.6966$. The calculation is shown in the following table where A is the event that "if Round R is played, you survive Round R" and B is the event that "you survive Round R" (this implies you survive from Round 1 to Round $R-1$).

Round (R)	$\Pr[A]$	$\Pr[B]$	Earned Point	Expectation
1	$52/52 = 1.0000$	1.0000	1	1.0000
2	$48/51 = 0.9412$	0.9412	1	0.9412
3	$44/50 = 0.8800$	0.8282	1	0.8282
4	$40/49 = 0.8163$	0.6761	1	0.6761
5	$36/48 = 9.7500$	0.5071	1	0.5071
6	$32/47 = 0.6809$	0.3452	1	0.3452
7	$28/46 = 0.6087$	0.2102	1	0.2102
8	$24/45 = 0.5333$	0.1121	1	0.1121
9	$20/44 = 0.4545$	0.0509	1	0.0509
10	$16/43 = 0.3721$	0.0190	1	0.0190
11	$12/42 = 0.2857$	0.0054	1	0.0054
12	$8/41 = 0.1951$	0.0011	1	0.0011
13	$4/40 = 0.1000$	0.0001	1	0.0001
14	$0/39 = 0.0000$	0.0000	1	0.0000
Sum	—	—	—	4.6966

We can also think about the probability that you will come out winning some homework points. The probability that you will survive round 4, $\Pr[B]$, is 0.6761 and the probabiliy that you will survive round 5 is 0.5071. By the end of round 4 and 5 you will have gained 4 and 5 homework points back, respectively. This means you have 67.61% chance of not losing the game, and just a little over 50% chance to get up to 1 whole extra homework point! □

9. Hosting a party

You are hosting a party, and as the host it is rude to talk to only some of your guests. In reality you should socialize with all your guests at least once. For the parts that are marked with **[VL]**, please complete your answer in the Virtual Lab skeleton.

(a) Your first strategy is to continually pick a person at random and go chat with him/her. Assume that in the party, there are n guests plus you. If you do the chatting strategy m times ($m \leq n$), what is the probability that you talk with everyone at most once?

This is exactly the birthday paradox, where different days in a year are replaced by different guests. We label each guest using number from 1 to n. Let $g_j \in \{1,2,\cdots,n\}$ be the guest you chatted with in the j-th conversation ($j = 1,\cdots,m$). Let A_t be the event that all g_j's up to the t-th trial are all different.

By the Product Rule,

$$
\begin{aligned}
&\Pr[\text{You talk with every guest at most once after chatting } m \text{ times}]\\
=&\Pr[A_m]\\
=&\Pr[g_m \neq g_{j,1\leq j\leq m-1}|A_{m-1}) \times \ Pr[A_{m-1}]\\
=&\Pr[g_m \neq g_{j,1\leq j\leq m-1}|A_{m-1}] \times \Pr[g_{m-1} \neq g_{j,1\leq j\leq m-2}|A_{m-2}] \times \ldots \times \Pr[g_3 \neq g_{j,1\leq j\leq 2}|A_2] \times \Pr[g_2 \neq g_1]\\
=&\frac{n-(m-1)}{n} \times \frac{n-(m-2)}{n} \times \cdots \times \frac{n-2}{n} \times \frac{n-1}{n}\\
=&\frac{(n-1)\times\cdots\times(n-m+1)}{n^{m-1}}\\
=&\frac{n!}{n^m(n-m)!}.
\end{aligned}
$$

Notice that $A_t = A_1 \cap A_2 \cap \ldots \cap A_{t-1} \cap A_t$.

An alternative solution is to select one of your n guests to chat with every time, so when you chat with your guests m times, there are a total of n^m possible outcomes. Among the n^m possible outcomes, there are $n(n-1)\cdots(n-(m-1))$ ways that all of the guests you have chatted with are different. Therefore the probability that you chat with everyone at most once after chatting m times is:

$$\frac{\frac{n!}{(n-m)!}}{n^m} = \frac{n!}{n^m(n-m)!}$$

(b) Now suppose that instead of randomly picking a single guest, you instead pick two people at random and chat with the one you've chatted with the least number of times (it does not matter how you break the ties). Now what is the chance that after doing this m times ($m \leq n$), you haven't talked to anyone more than once? (Assume $n \geq 2$.)

Based our strategy, notice that for $m = 1,2$, you will always talk to a guest who you have not been chatted with. So

$$\Pr[A_1] = \Pr[A_2] = 1.$$

For $m > 2$, again we have:

$$
\begin{aligned}
&\Pr[A_m]\\
=&\Pr[g_2 \neq g_1] \times \Pr[g_3 \neq g_{j,1\leq j\leq 2}|A_2] \times \ldots \times \Pr[g_i \neq g_{j,1\leq j\leq i-1}|A_{i-1}] \times \cdots \times \Pr[g_m \neq g_{j,1\leq j\leq m-1}|A_{m-1}],
\end{aligned}
\tag{1}
$$

where $\Pr[g_i \neq g_{j,1\leq j\leq i-1}|A_{i-1}]$ denotes the probability of chatting with a new guest in your i-th conversation given that you have chatted with $i-1$ different guests in your previous $i-1$ conversations.

Now we want to evaluate $\Pr[g_i \neq g_{j,1\leq j\leq i-1}|A_{i-1}]$. At your i-th conversation, you will randomly choose two guests, and there will be three cases:

- You have not chatted with both of the chosen guests. In this case, no matter who you choose to chat with, you will chat with a guest whom you haven't chatted with so far.
- You have chatted with one of the two guests chosen but not the other one. Based on the strategy you will again chat with a guest whom you haven't chatted with so far.
- You have chatted with both guests chosen. This is the only case that causes a problem. We will definitely have a guest whom you have chatted with more than once after this. This case happens with probability $\frac{\binom{i-1}{2}}{\binom{n}{2}}$.

Therefore the probability that you will chat with a new guest in your i-th conversation is $1 - \frac{\binom{i-1}{2}}{\binom{n}{2}}$, i.e.,

$$\Pr[g_i \neq g_{j,1\leq j\leq i-1} | A_{i-1}] = 1 - \frac{\binom{i-1}{2}}{\binom{n}{2}}.$$

Based on (1),

$$\begin{aligned}\Pr[A_m] &= 1 \times \left(1 - \frac{\binom{2}{2}}{\binom{n}{2}}\right) \times \left(1 - \frac{\binom{i-1}{2}}{\binom{n}{2}}\right) \times \ldots \times \left(1 - \frac{\binom{m-1}{2}}{\binom{n}{2}}\right) \\ &= \prod_{i=3}^{m} \left(1 - \frac{\binom{i-1}{2}}{\binom{n}{2}}\right) = \prod_{i=3}^{m} \left(1 - \frac{(i-1)(i-2)}{n(n-1)}\right) \quad \text{for } m \geq 3.\end{aligned}$$

(c) Suppose you use the first strategy. What is the expected number of times you should chat with people before having chatted with every guest at least once?

It is a Coupon Collector's Problem, where coupons are replaced by guests. The expected number of conversations needed to talk n guests at least once is

$$n \sum_{i=1}^{n} \frac{1}{i} \approx n(\ln n + \gamma),$$

where $\gamma = 0.5772$.

(d) Suppose you use the second strategy. What is the expected number of times you should chat with people before having chatted with every guest at least once? (Assume $n \geq 2$.)

Let X_i be the random variables denoting the additional conversations needed before chatting with the i-th new guest given that you have chatted with $i-1$ different guests. Let Y be the number of conversations needed before having chatted with every guest at least once. Then we have:

$$Y = X_1 + X_2 + \cdots + X_n.$$

For $i = 1, 2$, $X_i = 1$ with probability 1, so $\mathbf{E}[X_1] = \mathbf{E}[X_2] = 1$.

For $3 \leq i \leq n$, X_i has the geometric distribution with parameter $p_i = 1 - \frac{\binom{i-1}{2}}{\binom{n}{2}}$. Therefore

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{1}{1 - \frac{(i-1)(i-2)}{n(n-1)}}.$$

Thus, the expected number of conversations is

$$\mathbf{E}[Y] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = \begin{cases} 2 & \text{if } n = 2 \\ 2 + \sum_{i=3}^{n} \frac{1}{1 - \frac{(i-1)(i-2)}{n(n-1)}} & \text{if } n \geq 3 \end{cases}.$$

People who got the above answer are enough for full credit.

Now let's see how the expected number of guests $\mathbf{E}[Y]$ grows. Note that

$$\frac{(i-1)(i-2)}{n(n-1)} = \frac{(\frac{i}{n} - \frac{1}{n})(\frac{i}{n} - \frac{2}{n})}{1(1 - \frac{1}{n})}.$$

When n is large, we approximate $\frac{(i-1)(i-2)}{n(n-1)} \approx \frac{(i-1)^2}{n^2}$ by ignoring $\frac{1}{n}$terms. Then

$$\begin{aligned} 2+\sum_{i=3}^{n} \frac{1}{1-\frac{(i-1)(i-2)}{n(n-1)}} &= \sum_{i=1}^{n} \frac{1}{1-\frac{(i-1)(i-2)}{n(n-1)}} \\ &\approx \sum_{i=1}^{n} \frac{1}{1-\frac{(i-1)^2}{n^2}} \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{1-\frac{(i-1)}{n}}+\frac{1}{1+\frac{(i-1)}{n}}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\frac{1}{1-\frac{(i-1)}{n}}+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{1+\frac{(i-1)}{n}} \\ &= \frac{1}{2}\left(\frac{n}{n}+\frac{n}{n-1}+\cdots\frac{n}{1}\right)+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{1+\frac{(i-1)}{n}}. \end{aligned}$$

Since $\left(\frac{n}{n}+\frac{n}{n-1}+\cdots\frac{n}{1}\right) \approx n(\ln n+\gamma)$,

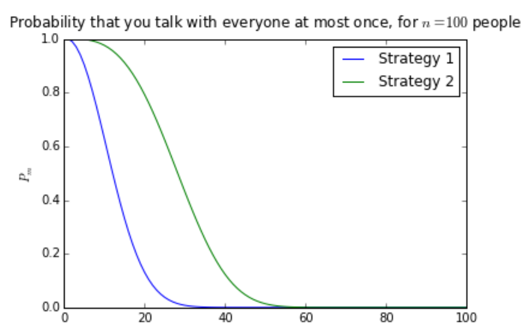
$$\mathbf{E}[Y] \approx \frac{1}{2}\left(\frac{n}{n}+\frac{n}{n-1}+\cdots\frac{n}{1}\right)+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{1+\frac{(i-1)}{n}} \approx \frac{1}{2}n(\ln n+\gamma)+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{1+\frac{(i-1)}{n}}.$$

Since $\frac{1}{2} \leq \frac{1}{1+\frac{(i-1)}{n}} \leq 1$,

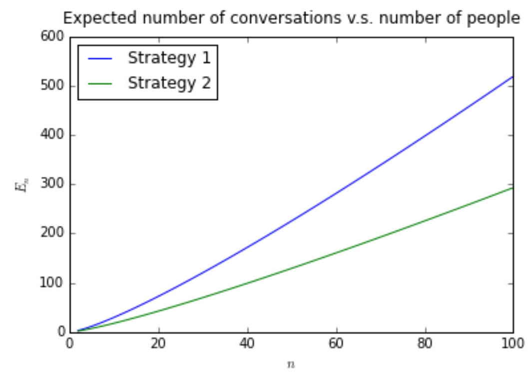
$$\frac{1}{2}n(\ln n+\gamma)+\frac{1}{4}n \leq \frac{1}{2}n(\ln n+\gamma)+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{1+\frac{(i-1)}{n}} \leq \frac{1}{2}n(\ln n+\gamma)+\frac{1}{2}n.$$

Therefore, the expected number of guests $\mathbf{E}[Y]$ is also growing in a rate of $O(n\ln n)$, but with a scalar $\frac{1}{2}$ compared to part (c). We can examine it by plotting them in part (f).

(e) **(*Extra Credit*) [VL]** Let $n = 100$. Plot the probabilities in (a) and (b) by varying m from 1 to n.

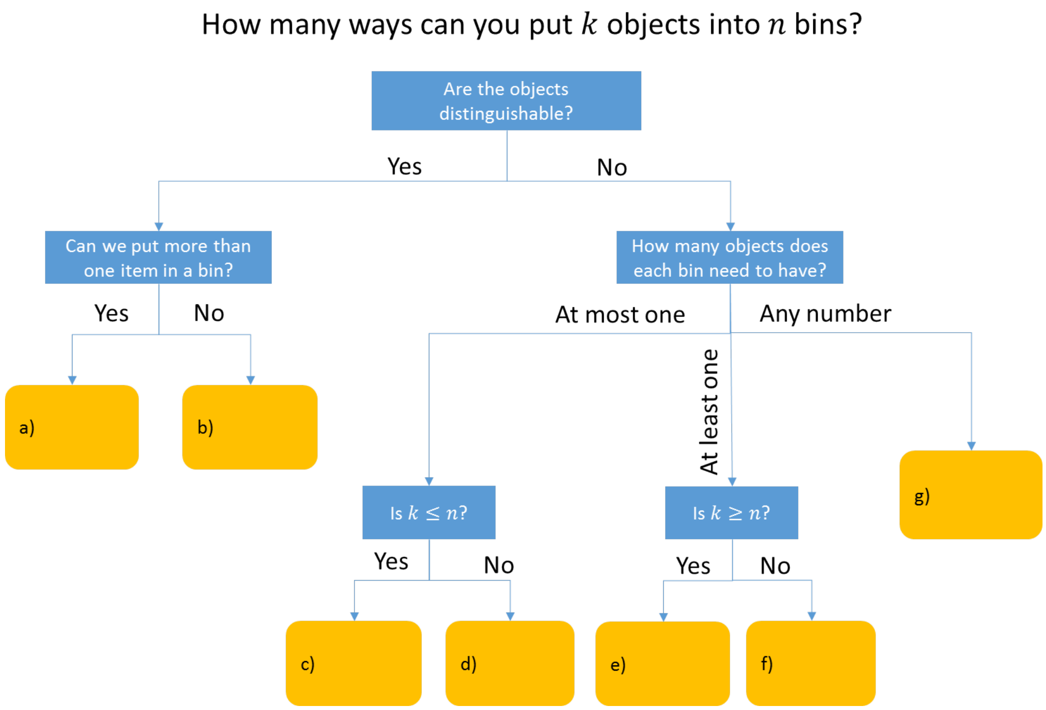


(f) **(*Extra Credit*) [VL]** Plot the expected numbers of times in (c) and (d) by varying n from 2 to 100.



10. ***(Extra Credit)* Fill in the Blanks!**

This is a flowchart for counting how many ways to put k objects into n bins. Fill in the blanks with the correct formula. Remember, each box corresponds to one subproblem.



(a) n^k. (Each ball has n choices of bin to go into.)

(b) $\frac{n!}{(n-k)!}$. (The first ball has n choices of bin, the second ball has $n-1$ choices of bin, etc.)

(c) $\binom{n}{k}$. (Choose k out of n bins to have the balls.)

(d) 0 (By pigeonhole principle, we will have at least one bin with more than one ball.)

(e) $\binom{k-1}{n-1}$. (Stars and bars. There are $k-1$ spaces between each ball and we select $n-1$ positions of them to insert bars to separate which balls should be in which bin.)
Summation forms such as

$$\sum_{n_1=1}^{b_1} \sum_{n_2=1}^{b_2} \cdots \sum_{n_{n-1}=1}^{b_{n-1}} \frac{k!}{n_1!n_2!\dots n_{n-1}!n_n!},$$

where $b_i = k - n + i - n_1 - n_2 - \ldots - n_{i-1}$ and $n_n = n - n_1 - n_2 - \ldots - n_{n-1}$, are acceptable, although the former answer is more preferable.

(f) 0 (There aren't enough balls to be in all bin.)

(g) $\binom{k+n-1}{n-1}$. (Think of a string of length $k+n-1$, we select $n-1$ positions of them to be bars, indicating each bin's boundary, and the rest to be balls.)
Again, summation forms such as

$$\sum_{n_1=0}^{b_1} \sum_{n_2=0}^{b_2} \cdots \sum_{n_{n-1}=0}^{b_{n-1}} \frac{k!}{n_1! n_2! \ldots n_{n-1}! n_n!},$$

where $b_i = k - n_1 - n_2 - \ldots - n_{i-1}$ and $n_n = n - n_1 - n_2 - \ldots - n_{n-1}$, are acceptable, although the former answer is more preferable.

□

11. Write your own problem

Write your own problem related to this week's material and solve it. You may still work in groups to brainstorm problems, but each student should submit a unique problem. What is the problem? How to formulate it? How to solve it? What is the solution?