Due: Saturday, 11/12, 4:00 PM
Grace period until Saturday, 11/12, 6:00 PM

Sundry

Before you start writing your final homework submission, state briefly how you worked on it. Who else did you work with? List names and email addresses. (In case of homework party, you can just describe the group.)

I worked on this homework with a couple people in thursday office hours. I also worked with **Christine Zhang (cyuzhang@berkeley.edu)** to complete this homework.

1 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 $\text{Var}(X)$. Your final answer should not contain any summations.

Solution: Let I_i be the indicator variable for the event that bin i has 1 ball. Now, we use:

$$E\left(\sum I_i\right) = \sum P(I_i)$$

to calculate the expected value. Computing the probability of any indicator variable I_i:

$$P(I_i) = \frac{1}{m}\left(\frac{m-1}{m}\right)^{n-1} n$$

We multiply by n at the end because there are n balls that we could choose from to go into bin i. Now, there are m bins, so therefore the total expected value is simply $mP(I_i)$, so therefore:

$$\sum E(I_i) = \sum P(I_i) = n\left(\frac{m-1}{m}\right)^{n-1}$$

Now we calculate $E(X^2)$. We use the relation

$$E(X^2) = \sum E\left(I_i^2\right) + 2\sum_{i<j} E\left(I_i I_j\right)$$

We know that $E(\sum I_i^2) = E(\sum I_i)$ which we found from part 1. To calculate the second term, note that the expected value is only nonzero when both $I_i = 1$ and $I_j = 1$. Therefore,

$$\sum_{i<j} E(I_i I_j) = \sum_{i<j} P(I_i I_j)$$
$$= \sum_{i<j} P(I_i = 1) \cdot P(I_j = 1 \mid I_i = 1)$$
$$= \frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1} \cdot \frac{n-1}{m-1}\left(\frac{m-2}{m-1}\right)^{n-2}$$

Therefore, we can compute the total $E(X^2)$:

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

And so therefore we can compute the variance:

$$\mathrm{Var}(X) = \frac{n}{m}\left(\frac{m-1}{m}\right)^{n-1}\left[m + 2\binom{m}{2}\frac{n-1}{m-1}\left(\frac{m-2}{m-1}\right)^{n-2}\right] - n^2\left(\frac{m-1}{m}\right)^{2n-2}$$

2 Will I Get My Package?

A delivery guy in some company is out delivering n packages to n customers, where n is a natural number greater than 1. Not only does he hand each customer a package uniformly at random from the remaining packages, he opens the package before delivering it with probability $1/2$. Let X be the number of customers who receive their own packages unopened.

(a) Compute the expectation $\mathbb{E}[X]$.

Solution: Let I_i be the indicator variable for the event that person i receives an unopened package. Then we know that

$$P(I_i) = \frac{1}{2} \cdot \frac{1}{n}$$

since the probability he selects a certain package is $\frac{1}{n}$, then the probaiblity the delivery guy doesn't open the package is $\frac{1}{2}$, so we multiply the two together. Since there are n people, then:

$$\begin{aligned} E\left(\sum I_i\right) &= \sum P(I_i) \\ &= n\frac{1}{2n} \\ &= \frac{1}{2} \end{aligned}$$

(b) Compute the variance $\mathrm{Var}(X)$.

Solution: Now we calculate $E(X^2)$. Using the same technique as problem 1, we use the fact that

$$\begin{aligned} E(X^2) &= E\left(\sum I_i\right) + 2E\left(\sum i < jI_iI_j\right) \\ &= \sum P(I_i) + 2\sum_{i<j} P(I_iI_j) \end{aligned}$$

We know from the first part that $\sum P(I_i) = \frac{1}{2}$, so all that remains is to calculate the latter half. We can do this quite easily:

$$\begin{aligned} E(I_iI_j) &= P(I_i = 1)P(I_j = 1 \mid I_i = 1) \\ &= \frac{1}{2} \cdot \frac{1}{n} \cdot \frac{1}{2} \cdot \frac{1}{n-1} \\ &= \frac{1}{4n(n-1)} \end{aligned}$$

So therefore:

$$E(x^2) = \frac{1}{2} + 2\binom{n}{2}\frac{1}{4n(n-1)}$$
$$= \frac{1}{2} + \frac{1}{4}$$

So now we can calculate Var(x) by combining this with our result from part (a):

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{2} + \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$$

3 Double-Check Your Intuition Again

(a) You roll a fair six-sided die and record the result X. You roll the die again and record the result Y.

   (i) What is $\text{cov}(X+Y, X-Y)$?

   Solution: The covariance of two random variables A and B can be written as $\text{cov}(A,B) = E(AB) - E(A)E(B)$. Therefore, we can rewrite this as:

   $$\text{cov}(X+Y, X-Y) = E((X+Y)(X-Y)) - E(X+Y)E(X-Y)$$
   $$= E(X^2) - E(Y^2) - E(X)^2 + E(Y)^2$$
   $$= \text{Var}(X) - \text{Var}(Y)$$

   Then, since X and Y are both independent dice rolls, then we have $\text{Var}(X) - \text{Var}(Y) = 0$ and so $\text{cov}(X+Y, X-Y) = 0$ as well.

   (ii) Prove that $X+Y$ and $X-Y$ are not independent.

   Solution: To prove that they are not independent then we need to show that $P(X+Y \mid X-Y) \neq P(X+Y)$. We can write the left hand side as:

   $$P(X+Y \mid X-Y) = \frac{P(X+Y \cap X-Y)}{P(X-Y)}$$

   But we know that $P(X+Y \cap X-Y) = 0$, since there are no two values X and Y such that $X+Y = X-Y$. However, the right hand side, $P(X+Y)$ is clearly nonzero (since it is a valid proability), so therefore both sides are not equal.

   For each of the problems below, if you think the answer is "yes" then provide a proof. If you think the answer is "no", then provide a counterexample.

(b) If X is a random variable and $\text{Var}(X) = 0$, then must X be a constant?

   Solution: Yes this is true. We know that $\text{Var}(x) = E[(x-\mu)^2] = 0$. Therefore, we can write:

   $$E[(x-\mu)^2] = 0$$
   $$\sum_i P_i(x_i - \mu)^2 = 0$$

   Since P_i must be nonzero for all i (otherwise it wouldn't be included in our sample space), so therefore $x_i - \mu = 0 \implies x_i = \mu$, which is a constant. Therefore, X must be a constant.

(c) If X is a random variable and c is a constant, then is $\text{Var}(cX) = c\,\text{Var}(X)$?

   Solution: No this is not true. Consider a roll of a six-sided dice, and choose $c = 2$. The variance of a dice roll is $35/12$ (calculated using $E(X) = 21/6$ and $E(X^2) = 91/6$). However, $\text{Var}(cX) = c^2\,\text{Var}(X) = \frac{182}{3}$, and so these two are not equal.

(d) If A and B are random variables with nonzero standard deviations and $\text{Corr}(A,B) = 0$, then are A and B independent?

Solution: If $\text{Corr}(A,B) = 0$, then this implies that $\text{cov}(A,B) = 0$. However, this does not imply that the two events are independent. Consider the following counterexample, as done in lecture:

Suppose we have 5 points on the cartesian plane: $(0,0), (1,0), (0,1), (-1,0), (0,-1)$, and select one of these points at random. Let X be the value of the x-coordinate and Y be the value of the y-coordinate. Then we know that $E(XY) = 0$, $E(X) = 0$ and $E(Y) = 0$ (from simple math), but these two events are clearly not independent, since:

$$P(X = 0 \mid Y = 0) = 1/3$$
$$P(X = 0) = 3/5$$

Adn since they're not equal, then X and Y are not independent events.

(e) If X and Y are not necessarily independent random variables, but $\text{Corr}(X,Y) = 0$, and X and Y have nonzero standard deviations, then is $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$?

Solution: We know that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X,Y)$ for any two events X and Y. Since we know that $\text{Corr}(X,Y) = 0$, then this implies that $\text{cov}(X,Y) = 0$, so therefore the first equation is true.

(f) If X and Y are random variables then is $\mathbb{E}[\max(X,Y)\min(X,Y)] = \mathbb{E}[XY]$?

Solution: This is true. The $\max(X,Y)$ will choose one of X or Y depending on which is greater, then $\min(X,Y)$ will simply choose the other. Since $E(XY) = E(YX)$, then these two equations are equivalent.

(g) If X and Y are independent random variables with nonzero standard deviations, then is

$$\text{Corr}(\max(X,Y), \min(X,Y)) = \text{Corr}(X,Y)?$$

Solution: This is not true. Consider the following counterexample:

Suppose we have a coin toss, and let I_1 represent the event that the first flip lands heads, and I_2 the second. Now we set a new random variable $X = \max(I_1, I_2)$ and $Y = \min(I_1, I_2)$. Because I_1 and I_2 are independent events, then $\text{cov}(I_1, I_2) = 0$, so $\text{Corr}(X,Y) = 0$. However, if we compute the left side:

$$\begin{aligned}
\text{cov}(\max(I_1, I_2), \min(I_1, I_2)) &= E(\max(I_1, I_2)\min(I_1, I_2)) \\
&= E(I_1 I_2) - E(\min(I_1, I_2))E(\max(I_1, I_2)) \\
&= \frac{1}{4} - \frac{1}{4} \cdot \frac{3}{4} = \frac{1}{16} \neq 0
\end{aligned}$$

And since this side is not equal to zero, then we've found a counterexample.

4 Fishy Computations

Assume for each part that the random variable can be modelled by a Poisson distribution.

(a) Suppose that on average, a fisherman catches 20 salmon per week. What is the probability that he will catch exactly 7 salmon this week?

Solution: This is a Poisson distribution with parameter $\lambda = 20$. Therefore:

$$P(X = 7) = \frac{20^7}{7!}e^{-20}$$

(b) Suppose that on average, you go to Fisherman's Wharf twice a year. What is the probability that you will go at most once in 2018?

Solution: The probability that you go at most once is equal to the sum of the probabilities that you don't go at all and that you go once. Mathematically:

$$P(k \leq 1) = P(k = 0) + P(k = 1) = e^{-2} + 2e^{-2} = 3e^{-2}$$

(c) Suppose that in March, on average, there are 5.7 boats that sail in Laguna Beach per day. What is the probability there will be *at least* 3 boats sailing throughout the *next two days* in Laguna?

Solution: If on averagre there are 5.7 boats per day, then there are also on average 11.4 boats per two days. This is still a Poisson distribution, but with a parameter of $\lambda = 11.4$. Therefore:

$$P(X > 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$$
$$= 1 - e^{-11.4} - 11.4e^{-11.4} - \frac{11.4^2}{2}e^{-11.4}$$
$$= 1 - e^{-11.4}\left(1 + 11.4 + \frac{11.4^2}{2}\right)$$

5 Geometric and Poisson

Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Poisson}(\lambda)$ be independent. random variables. Compute $\mathbb{P}[X > Y]$. Your final answer should not have summations.

Hint: Use the total probability rule.

Solution: Since Y is a Poisson distribution, the probability that $Y = k$ is:

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Now since X is a geometric distribution, we know that the probability that $X > k$ is:

$$
\begin{aligned}
P(X > k) &= p(1-p)^k + p(1-p)^{k+1} + \cdots \\
&= p(1-p)^k \left[1 + (1-p) + (1-p)^2 + \cdots \right] \\
&= p(1-p)^k \left[\frac{1}{1 - (1-p)} \right] \\
&= (1-p)^k
\end{aligned}
$$

Therefore, the probability that $X > Y$ is:

$$
\begin{aligned}
P(X > Y) &= \sum_{k=0}^{\infty} (1-p)^k \cdot \frac{\lambda^k}{k!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{[(1-p)\lambda]^k}{k!} \\
&= e^{\lambda} e^{(1-p)\lambda} \\
&= e^{-\lambda p}
\end{aligned}
$$

6 Poisson Coupling

(a) Let X, Y be discrete random variables taking values in $\mathbb{N}$. A common way to measure the "distance" between two probability distributions is known as the total variation norm, and it is given by

$$d(X,Y) = \frac{1}{2} \sum_{k=0}^{\infty} |\mathbb{P}[X=k] - \mathbb{P}[Y=k]|.$$

Show that

$$d(X,Y) \leq \mathbb{P}[X \neq Y]. \tag{1}$$

[*Hint*: Use the Law of Total Probability to split up the events according to $\{X = Y\}$ and $\{X \neq Y\}$. Also, the inequality $|a - b| \leq a + b$ might be helpful.]

Solution: We use the expression given in the problem statement, and turn it into an expression involving $X = Y$ and $X \neq Y$:

$$d(X,Y) = \frac{1}{2} \sum_{k=0}^{\infty} |P(X=k \cap X = Y + P(X=k \cap X \neq Y) - P(Y=k \cap X = Y) - P(Y=k \cap X \neq Y))|$$

Now since $P(X=k \cap X = Y) - P(Y=k \cap X = Y) = 0$ by the law of total probability, so therefore we can cancel these two terms. Then, we use the inequality given in the problem statement $|a - b| \leq a + b$:

$$\frac{1}{2} \sum_{k=0}^{\infty} |P(X=k \cap X \neq Y) - P(Y=k \cap X \neq Y)| \leq \frac{1}{2} \sum_{k=0}^{\infty} P(X=k \cap X \neq Y) + P(Y=k \cap X \neq Y)$$

$$\leq \frac{1}{2}(P(X \neq Y) + P(X \neq Y))$$

$$\leq P(X \neq Y)$$

And so therefore:

$$d(X,Y) \leq P(X \neq Y)$$

as desired.

(b) Show that if X_i, Y_i, $i \in \mathbb{Z}_+$ are discrete random variables taking values in $\mathbb{N}$, then $\mathbb{P}[\sum_{i=1}^{n} X_i \neq \sum_{i=1}^{n} Y_i] \leq \sum_{i=1}^{n} \mathbb{P}[X_i \neq Y_i]$. [*Hint*: Maybe try the Union Bound.]

Solution: Using the Union bound:

$$P\left(\bigcup_{i=1}^{n} X_i \neq Y_i\right) \leq \sum_{i=1}^{n} P(X_i \neq Y_i)$$

Now we aim to show:

$$P\left(\bigcup_{i=1}^{n} X_i \neq Y_i\right) \geq P\left(\sum X_i \neq \sum Y_i\right)$$

If we think about this intuitivley: if $X_i \neq Y_i$, it is still possible that $\sum X_i = \sum Y_i$, such as in the case $X_i = \{1,4\}$ and $Y_i = \{2,3\}$. However, if their sums are not equal, then it is necessarily true that for some i we have $X_i \neq Y_i$. Therefore, the space of possibilites for the left hand side is much larger than that of the right hand side, since there are instances where the left hand side is satisfied but the right is not.

Notice that the LHS of (1) only depends on the *marginal* distributions of X and Y, whereas the RHS depends on the *joint* distribution of X and Y. This leads us to the idea that we can find a good bound for $d(X,Y)$ by choosing a special joint distribution for (X,Y) which makes $\mathbb{P}[X \neq Y]$ small.

We will now introduce a coupling argument which shows that the distribution of the sum of independent Bernoulli random variables with parameters p_i, $i = 1,\ldots,n$, is close to a Poisson distribution with parameter $\lambda = p_1 + \cdots + p_n$.

(c) Let (X_i, Y_i) and (X_i, Y_j) be independent for $i \neq j$, but for each i, X_i and Y_i are *coupled*, meaning that they have the following discrete distribution:

$$\mathbb{P}[X_i = 0, Y_i = 0] = 1 - p_i,$$
$$\mathbb{P}[X_i = 1, Y_i = y] = \frac{e^{-p_i} p_i^y}{y!}, \qquad\qquad y = 1, 2, \ldots,$$
$$\mathbb{P}[X_i = 1, Y_i = 0] = e^{-p_i} - (1 - p_i),$$
$$\mathbb{P}[X_i = x, Y_i = y] = 0, \qquad\qquad \text{otherwise.}$$

Recall that all valid distributions satisfy two important properties. Argue that this distribution is a valid joint distribution.

Solution: To show that it is a valid distribution, we need to show that the $P(X = k)$ and $P(Y = k)$ is always nonnegative, and they sum to 1. Summing all the probabilities:

$$\sum P = (1 - p_i) + \sum_{y=1}^{\infty} \frac{p_i^y}{y!} + e^{-p_i} - (1 - p_i)$$
$$= e^{-p_i}\left(1 + \sum_{y=1}^{\infty} \frac{p_i^y}{y!}\right)$$
$$= e^{-p_i}(1 + e^{p_i} - 1)$$
$$= 1$$

Summing over Y_i is the same computation since we also sum over all the possible given terms, so it also sums to 1. To show that all terms are nonnegative, we go through them one by one. First:

$$P(X_i = 0, Y_i = 0) = 1 - p_i$$

Since $0 \le p_i \le 1$, then this probability is always nonnegative. For the second term:

$$P(X_i = 1, Y_i = y) = \frac{e^{-p_i} p_i^y}{y!}$$

We know that all three terms here are nonnegative, so therefore this term is also nonnegative. Now for the last term:

$$P(X_i = 1, Y_i = 0) = e^{-p_i} - (1 - p_i) = e^{-p_i} + p_i - 1$$

Graphically (using Desmos), we know that $e^{-p_i} + p_i$ has a minimum value of 1, so therefore this value is also nonnegative.

(d) Show that X_i has the Bernoulli distribution with probability p_i.

Solution: The probability that $X_i = 0$ is given as $1 - p_i$. then we calculate the probability that $X_i = 1$:

$$\begin{aligned}
P(X_i = 1) &= \sum_{y=1}^{\infty} \frac{e^{-p_i} p_i^y}{y!} + e^{-p_i} - (1 - p_i) \\
&= e^{-p_i}(1 + e^{p_i} - 1 - (1 - p_i)) \\
&= p_i
\end{aligned}$$

which is exaclty what we expect for a Bernoulli distribution.

(e) Show that Y_i has the Poisson distribution with parameter $\lambda = p_i$.

Solution: The probability that $Y_i = 0$ is $P(Y_i = 0) = (1 - p_i) + e^{-p_i} - (1 - p_i) = e^{p_i}$, which is what we expect for $P(Y = 0)$ if Y is a Poisson distribution. Then, the probability that $Y = y$ is given by:

$$P(Y = y) = \frac{e^{-p_i} p_i^y}{y!}$$

which is exactly the form for a Poisson distribution for $Y_i = y$. Therefore, the combination of these two probabilities gives us a Poisson distribution.

(f) Show that $\mathbb{P}[X_i \neq Y_i] \leq p_i^2$.

Solution: By the law of total probability, we know that $P(X \neq Y) = 1 - P(X = Y)$, so therefore:

$$\begin{aligned} p_i^2 &\geq 1 - P(X_i = Y_i) \\ &\geq 1 - P(X_i = 0, Y_i = 0) - P(X_i = 1, Y_i = 1) \\ &\geq 1 - (1 - p_i) - p_i e^{-p_i} \\ &\geq p_i(1 - e^{-p_i}) \end{aligned}$$

and so our inequality simplifies to:

$$1 - e^{-p_i} \leq p_i$$

and we know from part (c) that this inequality is true, and so we're done.

(g) Finally, show that $d(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i) \leq \sum_{i=1}^{n} p_i^2$.

Solution: We know from part (a) that:

$$d\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right) \leq P(X_i \neq Y_i)$$

And we know from the previous part we know that $P(X_i \neq Y_i) \leq p_i^2$, so therefore this equation then reads:

$$d\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right) \leq P(X_i \neq Y_i) \leq p_i^2$$

So therefore

$$d\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right) \leq \sum_{i=1}^{n} p_i^2$$

automatically follows.