Problem Set 11.5 Solutions

December 8, 2000

You do not have to turn in the solutions to this Problem Set. It is designed to help you study for the final.

Self-study: All about variance, from Section 2 of Lecture Notes 25.

Reading: Lecture notes 25 and 26; Rosen, Sections 4.4, 4.5,;
(Optional) Ross Sections 4.6, 8.2, 8.5.

Final Exam: Covers the entire course, with more emphasis on the last 8 lectures. The final will be held Thursday, December 21, 1:30 p.m. to 4:30 p.m. in Rockwell.

Problem 1

(a) Let X be a random variable whose value is an observation drawn uniformly at random from the $\{i \in \mathbb{Z} \mid -n \leq i \leq n\}$. Let $Y = X^2$. Show that $Ex[XY] = Ex[X]Ex[Y]$. Are X and Y independent?

Solution:

Since Y is determined by X, there are many cases indicating where X and Y are not independent, e.g.,

$$\Pr(Y = 0) = 1/(2n + 1), \text{ but } \Pr(Y = 0 \mid X = 0) = 1.$$

To see that $Ex[XY] = Ex[X]Ex[Y]$, note that $Ex[X] = 0$ and $Ex[XY] = Ex[X^3] = 0$ since X and X^3 are both symmetrical.

(b) Show that in general for any two random variables X and Y that if $Ex[XY] = Ex[X]Ex[Y]$ then $var(X+Y) = var(X)+var(Y)$. In particular, this shows that it is possible for nonindependent variables to have variances that exhibit linearity.

Solution:

We have that $var(X + Y) = Ex((X + Y)^2) - (Ex(X + Y))^2 = Ex(X^2 + 2XY + Y^2) - (Ex(X) + Ex(Y))^2 = Ex(X^2) - (Ex(X))^2 + Ex(Y^2) - (Ex(Y))^2 + 2Ex(XY) - 2Ex(X)Ex(Y) = var(X) + var(Y)$.

Problem 2 Suppose you are playing the game "Hearts" with three of your friends. In Hearts, all the cards are dealt to the players, in this case the four of you will each have 13 cards.

(a) What is the expectation and variance of the number of hearts that you will have in your hand?

Solution:

Let H be the number of hearts in your hand, and let X_i be the indicator variable that is 1 if card i is a heart and is 0 otherwise, so $H = \sum_{1 \le i \le 13} X_i$. Because X_i is an indicator variable, its expectation is the probability that card i is a heart, namely, $1/4$. So

$$\mathrm{E}[H] = \sum_{1}^{13} 1/4 = 13/4.$$

To compute variance, we use the definition:

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{13} X_i\right] &= \mathrm{E}\left[(\sum_{i=1}^{13} X_i)^2\right] - \mathrm{E}\left[\sum_{i=1}^{13} X_i\right]^2 \\
&= \mathrm{E}\left[(\sum_{i=1}^{13} X_i)^2\right] - (13/4)^2 \\
&= \mathrm{E}[X_1^2 + X_1X_2 + X_1X_3 + \ldots + X_1X_{13} + \\
&\quad X_2X_1 + X_2^2 + X_2X_3 + \ldots + X_2X_{13} + \\
&\quad \vdots \\
&\quad X_{13}X_1 + X_{13}X_2 + \ldots + X_{13}^2] - (13/4)^2 \\
&= \mathrm{E}\left[X_1^2\right] + \mathrm{E}[X_1X_2] + \mathrm{E}[X_1X_3] + \ldots + \mathrm{E}[X_1X_{13}] + \\
&\quad \mathrm{E}[X_2X_1] + \mathrm{E}\left[X_2^2\right] + \mathrm{E}[X_2X_3] + \ldots + \mathrm{E}[X_2X_{13}] + \\
&\quad \vdots \\
&\quad \mathrm{E}[X_{13}X_1] + \mathrm{E}[X_{13}X_2] + \ldots + \mathrm{E}\left[X_{13}^2\right] - (13/4)^2.
\end{aligned}$$

Now we note that $\mathrm{E}\left[X_i^2\right] = \mathrm{E}[X_i]$ since X_i is an indicator variable, so $\mathrm{E}\left[X_i^2\right] = 1/4$. Also, when $i \neq j$, $X_iX_j = 1$ if cards i and j are both hearts, and 0 otherwise. The probability that both cards are hearts is $(1/4)(12/51)$, and thus $\mathrm{E}[X_iX_j] = (1/4)(12/51) = 3/51$. In the summation above, there are 13 $\mathrm{E}\left[X_i^2\right]$ terms, and $2\binom{13}{2} = 156$ terms of the form $\mathrm{E}[X_iX_j]$. Thus,

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{13} X_i\right] &= 13 \cdot \frac{1}{4} + 156 \cdot \frac{3}{51} - (\frac{13}{4})^2 \\
&\approx 1.863970588.
\end{aligned}$$

(b) What is the expectation and variance of the number of suits in your hand?

Solution:

Similar to the first part: let N denote the number of suits in a hand. Let X_s be the indicator variable which is 1 if there is a spade in the hand and 0 otherwise. Define X_h, X_d, X_c analogously for the three remaining suits: hearts, diamonds, and clubs. Then $N = X_s + X_h + X_d + X_c$.

The expectation of each of the indicator variables is the probability that a hand contains the corresponding suit. Thus, for $i \in \{s, h, d, c\}$

$$\mathrm{E}\left[X_i\right] = 1 - \frac{\binom{52-13}{13}}{\binom{52}{13}} = 0.987209,$$

because $\binom{52-13}{13}$ is the number of hands that are missing the particular suit, and $\binom{52}{13}$ is the total number of hands.

$$\mathrm{E}\left[N\right] = \mathrm{E}\left[X_s + X_h + X_d + X_c\right] = \mathrm{E}\left[X_s\right] + \mathrm{E}\left[X_h\right] + \mathrm{E}\left[X_d\right] + \mathrm{E}\left[X_c\right] = 4 \cdot 0.987209 = 3.94884.$$

To compute variance:

$$\begin{aligned} \mathrm{Var}\left[N\right] &= \Pr(N=1)(1-\mathrm{E}\left[N\right])^2 + \Pr(N=2)(2-\mathrm{E}\left[N\right])^2 + \\ &\qquad \Pr(N=3)(3-\mathrm{E}\left[N\right])^2 + \Pr(N=4)(4-\mathrm{E}\left[N\right])^2. \end{aligned}$$

Now we must calculate each of the probabilities. First,

$$\Pr(N=1) = \frac{4}{\binom{52}{13}} = 6.299 \times 10^{-12},$$

because there are four possible hands of all one suit out of $\binom{52}{13}$ possible hands of 13 cards.

$$\Pr(N=2) = \frac{\binom{4}{2}\left[\binom{26}{13} - 2\right]}{\binom{52}{13}} = 9.827 \times 10^{-5},$$

because there are $\binom{4}{2}$ ways to choose which two suits. There are $\binom{26}{13}$ ways to choose 13 cards from the 26 possible cards of those two suits. Two of those $\binom{26}{13}$ hands actually contain only one suit. Thus $\binom{4}{2}(\binom{26}{13} - 2)$ is the number of hands with two suits.

$$\Pr(N=3) = \frac{\binom{4}{3}\left[\binom{39}{13} - \binom{3}{2}\binom{26}{13} + 3\right]}{\binom{52}{13}} = 0.05097,$$

because there are $\binom{4}{3}$ ways to choose the three suits. There are $\binom{39}{13}$ ways to choose 13 cards from the 39 cards of those three suits. But that also includes hands with two suits and one suit. So, using Inclusion-Exclusion, we subtract the $\binom{3}{2}\binom{26}{13}$ ways to choose two suits from the three suits and then choose choose 13 cards from the 26 cards of those suits. Next, we add back the three ways to choose hands of one suit among the three. Finally, we divide by the $\binom{52}{13}$ ways to choose a possible hand of 13.

Of course,

$$\Pr(N=4) = 1 - (\Pr(N=1) + \Pr(N=2) + \Pr(N=3)) = 0.9489.$$

This allows us to complete the calculation of the variance:

$$\mathrm{Var}\left[N\right] = 0.04874.$$

Problem 3 Psychological diagnosis of sociopaths is notoriously unreliable. For example, a random psychologist is as likely to diagnose a normal patient as sociopathic as he is to recognize that the patient is normal. A random psychologist is actually slightly *less* likely to correctly diagnose a real sociopath than not—sociopaths are really good liars.

(a) In these circumstances it might seem there was no reliable way to diagnose sociopaths, but in theory there is a way—if we have a large enough population of psychologists who reach their judgements independently. Explain how to obtain a reliable diagnosis in this case. Briefly explain why your method will be reliable, citing appropriate results from class. A qualitative answer is all that's required—you need not come up with any numerical estimates or growth rates.

Solution:

Get a reliable test by repeatedly asking randomly chosen psychologists to do the diagnosis and computing the fraction, f, of "normal" diagnoses. Let $s > 1/2$ be the probability of a "normal" diagnosis on a sociopathic patient. Diagnose normal if f is closer to $1/2$ than to s.

By diagnosing independently enough times, there is small probability that f differs from the expected value of a diagnosis by very much (the expected value is either $1/2$ or s depending on whether or not the patient is normal—we don't know which). By the Law of Large Numbers, by Chebyshev's bound, or by the estimation of the binomial distribution as used in the Gallup Poll example in Lecture Notes, if we diagnose enough times, we can make this probability as small as desired, so the probability of a correct diagnosis can be as close to 1 as desired.

(b) Is it plausible to assume that psychologists will make their diagnoses independently? Briefly explain.

Solution:

Not really. It figures that some sociopaths would be particularly good liars and be able to fool most psychologists. Conversely, there figure to be normal people whose anxious or defensive manner makes a sociopathic diagnosis by many psychologists more likely. Also, there may be some specially expert psychologists whose diagnoses are more reliable than average; this implies if one expert diagnoses "normal," the diagnosis is likely to be correct, making it more likely that another expert will also correctly diagnose "normal," so their diagnoses will not be independent of each other.

Problem 4 Three hundred students take the 6.042 final exam. The mean on the exam is 90 and the lowest score was 30. You have no other information about the students and the exam, e.g., you should not assume that the final is worth 100 points.

(a) State the best possible upper bound on the number of students who scored at least 180.

Solution:

Let R be the score. Apply Markov's Bound to $R - 30$:

$$\Pr(R \geq 180) = \Pr(R - 30 \geq 150) \leq \mathrm{E}[R - 30]/150 = 60/150 = 2/5,$$

so at most $(2/5)300 = \mathbf{120}$ students score ≥ 180.

(b) Give an example set of scores which achieve your bound.

Solution:
Of the 300 students, 120 score 180 and 180 score 30. The mean is

$$(120 \cdot 180 + 180 \cdot 30)/300 = 90,$$

as required.

Problem 5 Suppose that n people have their hats returned at random. Let $X_i = 1$ if the ith person gets his or her own hat back and 0 otherwise. Let $S_n = \sum_{i=1}^{n} X_i$, so S_n is the total number of people who get their own hat back. Show that

(a) $\mathrm{E}\left[X_i^2\right] = 1/n$.

Solution:
$X_i = 1$ with probability $1/n$ and 0 otherwise. Thus $X_i^2 = 1$ with probability $1/n$ and 0 otherwise. So $\mathrm{E}\left[X_i^2\right] = 1/n$.

(b) $\mathrm{E}\left[X_i X_j\right] = 1/n(n-1)$ for $i \neq j$.

Solution:
The probability that X_i and X_j are both 1 is $1/n \cdot 1/(n-1) = 1/n(n-1)$. Thus $X_i X_j = 1$ with probability $1/n(n-1)$, and is zero otherwise. So $\mathrm{E}\left[X_i X_j\right] = 1/n(n-1)$.

(c) $\mathrm{E}\left[S_n^2\right] = 2$. *Hint*: Use (a) and (b).

Solution:

$$\begin{aligned}
\mathrm{E}\left[S_n^2\right] &= \sum_i \mathrm{E}\left[X_i^2\right] + \sum_i \sum_{j \neq i} \mathrm{E}\left[X_i X_j\right] \\
&= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} \\
&= 2.
\end{aligned}$$

(d) $\mathrm{Var}\left[S_n\right] = 1$.

Solution:

$$\begin{aligned}
\mathrm{Var}\left[S_n\right] &= \mathrm{E}\left[S_n^2\right] - \mathrm{E}^2\left[S_n\right] \\
&= 2 - (n(1/n))^2 \\
&= 2 - 1 \\
&= 1.
\end{aligned}$$

(e) Using Chebyshev's Inequality, show that $\Pr(S_n \geq 11) \leq .01$ for any $n \geq 11$.

Solution:

$$\begin{aligned}
\Pr(S_n \geq 11) &= \Pr(S_n - \mathrm{E}[S_n] \geq 11 - \mathrm{E}[S_n]) \\
&= \Pr(S_n - \mathrm{E}[S_n] \geq 10) \\
&\leq \frac{\mathrm{Var}[S_n]}{10^2} = .01
\end{aligned}$$

Note that the X_i's are Bernoulli variables but are *not* independent, so S_n does not have a Bernoulli distribution and the Bernoulli estimates from Lecture Notes 21 do not apply.

Problem 6

(a) Knowing only that the average graduating MIT student's total number of credits is 200, find a tight upper bound for the fraction of MIT students graduating with at least 235 credits. (Ignore the fact that students who graduate do so after at most 700 years; that is, assume that there is no theoretical upper bound on the possible number of credits a student may earn.)

Solution:

We use Markov's inequality $\Pr(X \geq c) \leq E[X]/c$, with $c = 235$, to obtain the upper bound $200/235 \approx .85$.

(b) Assuming in addition only that the standard deviation of the total credits per graduating student is 7, give a tight bound on the fraction of students who can graduate with at least 235 credits.

Solution:

We use the one-way Chebychev inequality $\Pr(X \geq E[X] + c\sigma_X) \leq \frac{1}{1+c^2}$ with $c = 5$ to get an upper bound of $1/26$.

(c) Different information about MIT credits is now provided by the Registrar. The standard deviation is not 7 after all. Instead, the registrar reports that 30% of past graduates are EECS majors and half of them had 235 or more credits. Only 1/7th of the non-EECS students graduated with that many credits. Use this information and the Chernoff bound to give a tight upper bound on the probability that among 1000 randomly chosen graduating students, at least 300 have 235 or more credits. (You may leave ln and e in your answer).

Solution:

We use the Chernoff bound. Among 1000 students, let N be the number with at least 235 credits. So $\mathrm{E}[N] = (1/7)700 + (1/2)300 = 250$ and

$$\begin{aligned}
\Pr(N \geq 300) &= \Pr(X \geq (1.2)250) \leq e^{-(\ln 1.2 + 1/(1.2) - 1)(1.2)250} = e^{-(\ln 1.2 - .1666)300} \\
&= e^{-(.182 - .1666)300} \leq e^{-4.6} \approx .01.
\end{aligned}$$

Problem 7 [[[Here's an idea; it needs polishing, solving, etc.]]]

The latest recount of the Florida votes has yielded exactly 3 million votes for each. The Florida Supreme Court has decided to allow repeated recounts of all the votes in Florida, until the difference between the two is at least 1000 votes.

It turns out that recounting of votes has the odd property that it can be modeled by a random walk. Specifically, each recount has probability .4 of yielding exactly 100 more votes for Bush than the previous count, and exactly .6 of giving 100 more votes for Gore. (Unlike the current election, in this scenario, the total number of votes after each recount is constant.)

(a) What is the probability of Bush eventually winning? Of Gore eventually winning? Of neither eventually winning (that is, recounts go on forever)?

Solution: Let P_n be the probability that Bush wins the election where n represents the net number of recounts won by Bush (for example, if Bush wins the first one and then loses the second and third, $n = -1$). Bush wins if he gets 1000 more votes (n=10) and loses if Gore gets 1000 more votes (n=-10). So, we can write the following recurrence for P_n.

$$P_n = \begin{cases} 0 & \text{if } n = -10 \\ 1 & \text{if } n = 10 \\ .4 * P_{n+1} + .6 * P_{n-1} & \text{if } -10 < n < 10 \end{cases}$$

This leads to the recurrence $.4P_{n+1} - P(n) + .6Pn - 1 = 0$, which has the characteristic equation $r^2 - 2.5 * r + 1.5 = 0$, where $P_{10} = 1$ and $P_{-10} = 0$. Using the quadratic formula, we find that the roots are r=1.5 and r=1. This gives us the solution:

$$P_n = A\left(\frac{3}{2}\right)^n + B(1)^n$$

Using the boundary conditions, $P_(10) = 1$ and $P_(- 10) = 0$, we can find the constants A and B. Substituting these two conditions into the general form gives:

$$\begin{aligned} 0 &= A * 1.5^{10} + B \\ 1 &= A * 1.5^{-10} + B \end{aligned}$$

Solving this system, we get $A = .0173$, $B = -3 * 10^-4$. Finally, solving for P_0, the probability that Bush eventually wins the election:

$$\begin{aligned} P_0 &= .0173 * (1.5)^0 + -3 * 10^{-4} * (1)^0 \\ &= .0170 \end{aligned}$$

So, the probability Bush wins is .017. By the analysis of theorem 3.1 in Lecture 26, the probability that no one ever wins is 0, and so the probability that Gore eventually wins is 1-.017 = .983

(b) What is the expected duration of the election (assume one recount per day)?

Solution:

By the analysis from Section 3 of Lecture 26, we have the following formula for E_n, the expected number of recounts of the election, given Bush is currently ahead by n recounts (note that n can range from -10 to 10).

$$E_n = \begin{cases} 0 & \text{if } n = -10 \\ 0 & \text{if } n = 10 \\ 1 + .4 * E_{n+1} + .6 * E_{n-1} & \text{if } -10 < n < 10 \end{cases}$$

Solving the homogeneous part of this recurrence gives the equation

$$\begin{aligned} E_n &= A\left(\frac{1-.4}{.4}\right)^n + B \\ &= A * 1.5^n + B \end{aligned}$$

Solving the particular part of the recurrences gives us $E_n = \frac{n}{1-(2*.4)}$. Adding the two parts together gives a complete solution of the form:

$$E_n = A(1.5)^n + B + \frac{n}{.2}$$

Using the boundary conditions $E_{-10} = E_{10} = 0$, we get the system of equations:

$$0 = A(1.5)^{10} + B + \frac{10}{.2}$$
$$0 = A(1.5)^{-10} + B + \frac{10}{.2}$$

This gives $A = 0$, $B = 50$. Finally, solving the equation given that $n = 0$ gives us the desired answer. $E_0 = 0 * (1.5)^n + 50 + \frac{0}{.2} = 50$ days until a winner is declared.

(To get a bit of intuition - which is always dangerous when doing probabilities! - we note that on average, Gore wins a recount 60 percent of the time and Bush wins one 40 percent of the time. So, if there are 10 recounts, say, Gore should win 6 of them, and Bush 4 of them, for a net gain of 2 recounts for Gore. Similarly, if there are 20 recounts, then Gore would win about 12 of them, and Bush would win about 8 of them, for a net gain of 4 for Gore. So, by this logic, if Gore wants a net gain of ten recounts, there should be 50 recounts total - since Gore would be expected to win about 30 of them, and Bush about 20 of them.)

(c) Gore has decided that if Bush wins according to this process, he (Gore) will throw in the towel. Bush, however, has decided that if Gore wins according to this process, he (Bush) will insist on continuing the recounts forever until he (Bush) has a 1000 vote margin. Now assume that Gore wins, according to this process.

i) How many recounts would be expected to occur until Bush wins?
ii) What is the probability that Bush eventually wins?