Problem Set 11 Solutions

Due: December 4, 7pm

Reading Assignment: Chapters 18, 19

Problem 1. [15 points]

In this problem, we will (hopefully) be making tons of money! Use your knowledge of probability and statistics to keep from going broke!

Suppose the stock market contains N types of stocks, which can be modelled by independent random variables. Suppose furthermore that the behavior of these stocks is modelled by a double-or-nothing coin flip. That is, stock S_i has half probability of doubling its value and half probability of going to 0. The stocks all cost a dollar, and you have N dollars. Say you only keep these stocks for one time-step (that is, at the end of this timestep, all stocks would have doubled in value or gone to 0).

(a) [3 pts] What is your expected amount of money if you spend all your money on one stock? Your variance?

Solution. The stock doubles on a coin flip, so your expected final amount is $.5(2N)+.5(0) = N$. Your variance is calculated as $E[(X-\mu)^2]$. This, when we take into account the probability distribution of the stock, is

$$1/2(2N-N)^2 + 1/2(0-N)^2 = N^2$$

■

(b) [3 pts] Suppose instead you diversified your purchases and bought N shares of all different stocks. What is your expected amount of money then? Your variance?

Solution. The amount of money you have in stocks is $X_1 + X_2 + \ldots + X_N$, where X_i is a random variable describing how much money you have in stock i. The amount of money you expect to have is

$$E[\sum_{i=1}^{N} X_i] = \sum_{i=1}^{N} E[X_i]$$

But $E[X_i] = 2 * 1/2 + 0 * 1/2 = 1$, so this sum turns out to be N again.

As for variance, recall that $Var(\sum X_i) = \sum Var(X_i)$ if the random variables X_i are independent (which they are in this case). The variance of a single X_i is

$$1/2 * (2-1)^2 + 1/2 * (0-1)^2 = 1$$

so the variance of your entire portfolio is N. ■

(c) [3 pts] The money that you have invested came from your financially conservative mother. As a result, your goals are much aligned with hers. Given this, which investment strategy should you take?

Solution. Your mother prefers to have less risk, and so would like the stock with less variance. This is the strategy associated with (b). ■

(d) [3 pts] Now instead say that you make money on rolls of dice. Specifically, you play a game where you roll a standard six-sided dice, and get paid an amount (in dollars) equal to the number that comes up. What is your expected payoff? What is the variance?

Solution. The expected payoff is $1/6(1+2+3+4+5+6) = 3.5$. The variance is

$$1/6(2.5^2 + 1.5^2 + .5^2 + .5^2 + 1.5^2 + 2.5^2) = 35/12$$

■

(e) [3 pts] We change the rules of the game so that your payoff is the cube of the number that comes up. In that case, what is your expected payoff? What is its variance?

Solution. Your expected earnings is $1/6(1+8+27+64+125+216) = 441/6$. To calculate variance, we can simplify by noting that $Var(X) = E[X^2] - E^2[X]$. So the variance is

$$1/6(1+64+27^2+64^2+125^2+216^2) - (441//6)^2 = 67171/6 - 194481/36 = 208545/36 \approx 5792$$

■

Problem 2. [8 points] Here are seven propositions:

$$\begin{array}{rcrcr}
x_1 & \vee & x_3 & \vee & \neg x_7 \\
\neg x_5 & \vee & x_6 & \vee & x_7 \\
x_2 & \vee & \neg x_4 & \vee & x_6 \\
\neg x_4 & \vee & x_5 & \vee & \neg x_7 \\
x_3 & \vee & \neg x_5 & \vee & \neg x_8 \\
x_9 & \vee & \neg x_8 & \vee & x_2 \\
\neg x_3 & \vee & x_9 & \vee & x_4
\end{array}$$

Note that:

1. Each proposition is the OR of three terms of the form x_i or the form $\neg x_i$.

2. The variables in the three terms in each proposition are all different.

Suppose that we assign true/false values to the variables $x_1, \ldots, x_9$ independently and with equal probability.

(a) [4 pts] What is the expected number of true propositions?

Solution. Each proposition is true unless all three of its terms are false. Thus, each proposition is true with probability:

$$1 - \left(\frac{1}{2}\right)^3 = \frac{7}{8}$$

Let T_i be an indicator for the event that the i-th proposition is true. Then the number of true propositions is $T_1 + \ldots + T_7$ and the expected number is:

$$\begin{aligned} \mathrm{E}[T_1 + \ldots + T_7] &= \mathrm{E}[T_1] + \ldots + \mathrm{E}[T_7] \\ &= 7/8 + \ldots + 7/8 \\ &= 49/8 = 6\frac{1}{8} \end{aligned}$$

■

(b) [4 pts] Use your answer to prove that there exists an assignment to the variables that makes *all* of the propositions true.

Solution. A random variable can not always be less than its expectation, so there must be some assignment such that:

$$T_1 + \ldots T_7 \geq 6\frac{1}{8}$$

This implies that $T_1 + \ldots + T_7 = 7$ for at least one outcome. This outcome is an assignment to the variables such that all of the propositions are true. ■

Problem 3. [10 points] We have two coins: one is a fair coin and the other is a coin that produces heads with probability 3/4. One of the two coins is picked, and this coin is tossed n times. Explain how to calculate the number of tosses to make us 95% confident which coin was chosen. You do not have to calculate the minimum value of n, though we'd be pleased if you did.

Solution. To guess which coin was picked, set a threshold t between 1/2 and3/4. If the proportion of heads is less than the threshold, guess it was the fair coin; otherwise, guess the biased coin. Let the random variable J be the number of heads in the first n flips. We need to flip the coin enough times so that $\Pr\{J/n \geq t\} \leq 0.05$ if the fair coin was picked, and $\Pr\{J/n \leq t\} \leq 0.05$ if the biased coin was picked. A natural threshold to choose is 5/8, exactly in the middle of 1/2 and 3/4.

For the fair coin, J has an $(n, 1/2)$-binomial distribution, so we need to choose n so that

$$\Pr\left\{J > \left(\frac{5}{8}\right) n\right\} \leq 0.05$$

which is equivalent to

$$CDF_J\left(\frac{5}{8}n\right) \geq 0.95 \tag{1}$$

For the biased coin, J has an $(n, 3/4)$-binomial distribution, so we need to choose n so that

$$\Pr\left\{J \leq \left(\frac{5}{8}\right) n\right\} \leq 0.05$$

which is equivalent to

$$CDF_J\left(\frac{5}{8}n\right) \leq 0.95 \tag{2}$$

We can now search for the minimum n that satisfies both (1) and (2), using one of the several ways we know to calculate or approximate the binomial cumulative distribution function.

■

Problem 4. [22 points]

Suppose n balls are thrown randomly into n boxes, so each ball lands in each box with uniform probability. Also, suppose the outcome of each throw is independent of all the other throws.

(a) [5 pts] Let X_i be an indicator random variable whose value is 1 if box i is empty and 0 otherwise. Write a simple closed form expression for the probability distribution of X_i. Are $X_1, X_2, \ldots, X_n$ independent random variables?

Solution. Box i is empty iff all n balls land in other boxes. The probability that a ball will land in another box in $(n-1)/n = 1-(1/n)$, and since the balls are thrown independently, we have

$$\Pr\left(X_i = 1\right) = \left(1 - \frac{1}{n}\right)^n. \tag{3}$$

The X_i's are not independent. For example,

$$\Pr\left(X_1 = X_2 = \cdots = X_n = 1\right) = 0 < \prod_{i=1}^{n} \Pr\left(X_i = 1\right).$$

■

(b) [2 pts] Find a constant, c, such that the expected number of empty boxes is asymptotically equal ($\sim$) to cn.

Solution. The number of empty boxes is the sum of the X_i's. So the expected number of empty boxes is the sum of the expectations of the X_i'. By (3), we now have

$$\operatorname{Ex}(\text{number of empty boxes}) = n \operatorname{Ex}(X_1) = n\left(1 - \frac{1}{n}\right)^n \sim n \cdot \frac{1}{e}$$

That is,

$$c = \frac{1}{e}$$

■

(c) [5 pts] Show that

$$\Pr(\text{at least } k \text{ balls fall in the first box}) \leq \binom{n}{k}\left(\frac{1}{n}\right)^k.$$

Solution. Let S be a set of k of the n balls, and let E_S be the event that each of these k balls falls in the first box. Since the probability that a ball lands in this box is $1/n$, and the throws are independent, we have

$$\Pr(E_S) = \left(\frac{1}{n}\right)^k. \tag{4}$$

The event that *at least* k balls land in the first box is the union of all the events E_S. There are $\binom{n}{k}$ subsets, S, of k balls, so by the Union Bound,

$$\Pr(\text{at least } k \text{ balls fall in the first box}) \leq \binom{n}{k} \cdot \Pr(E_S).$$

Using the value for $\Pr(E_S)$ from (4) in the preceding inequality yields the required bound.
■

(d) [5 pts] Let R be the maximum of the numbers of balls that land in each of the boxes. Conclude from the previous parts that

$$\Pr\{R \geq k\} \leq \frac{n}{k!}.$$

Solution. Note that $R \geq k$ exactly when some box has at least k balls. Since the bound on the probability of at least k balls in the first box applies just as well to any box, we can apply the Union Bound to having at least k balls in at least one of the n boxes:

$$\Pr(R \geq k) \leq n \Pr(\text{at least } k \text{ balls fall in the first box}).$$

So from the previous problem part, we have

$$\begin{aligned}
\Pr(R \geq k) &\leq n\binom{n}{k}\left(\frac{1}{n}\right)^k \\
&= n\left(\frac{n(n-1)\cdots(n-k+1)}{k!\,n^k}\right) \\
&= \frac{n}{k!}\left(\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right) \\
&\leq \frac{n}{k!}
\end{aligned}$$

■

(e) [5 pts] Conclude that

$$\lim_{n\to\infty} \Pr\{R \geq n^\epsilon\} = 0$$

for all $\epsilon > 0$.

Solution. Using Stirling's formula, and the upper bound from the previous part, we have

$$\Pr\{R \geq k\} \leq \frac{n}{k!} \sim \frac{n}{\sqrt{2\pi k}(k/e)^k} \leq \frac{n}{(k/e)^k} = \frac{ne^k}{k^k} = \frac{e^{k+\ln n}}{e^{k\ln k}}.$$

Now let $k = n^\epsilon$. Then the exponent of e in the numerator above is $n^\epsilon + \ln n$, and the exponent of e in the denominator is $n^\epsilon \ln n^\epsilon$. Since

$$n^\epsilon + \ln n = o(n^\epsilon \ln n^\epsilon),$$

we conclude

$$\Pr\{R \geq n^\epsilon\} \leq \frac{e^{n^\epsilon+\ln n}}{e^{n^\epsilon \ln n^\epsilon}} \to 0$$

as n approaches ∞. ■

Problem 5. [13 points] The goal of this problem will be to study the expectations of quotients of positive random variables.

Throughout, you may assume the following inequality: for all $\lambda \in (0,1)$, and all positive reals x and y,

$$\frac{1}{\lambda x + (1-\lambda)y} \geq \frac{\lambda}{x} + \frac{1-\lambda}{y}.$$

(This property has a name: the function $f(x) = \frac{1}{x}$ is *convex*.)

(a) [10 pts] Let X be a positive random variable with finitely many outcomes. Prove that

$$E\left[\frac{1}{X}\right] \geq \frac{1}{E[X]}.$$

(*Hint: Try induction on the number of outcomes of X.*)

Solution. We proceed by induction on n the number of outcomes of X. The base case is easily verified: if X has only one outcome x, then

$$E\left[\frac{1}{X}\right] = \frac{1}{x} = \frac{1}{E[X]}.$$

Suppose then that the result is true for all positive random variables on at most n outcomes, and let X have $n+1$ outcomes $x_1, \dots, x_{n+1}$ with probabilities $p_1, \dots, p_{n+1}$. Let Y be the random variable consisting of only the first n outcomes $x_1, \dots, x_n$, with probabilities

$$\frac{p_1}{1-p_{n+1}}, \dots, \frac{p_n}{1-p_{n+1}}$$

rescaled in order to add up to 1. Applying the inductive hypothesis to Y,

$$\frac{1}{\sum_{i=1}^n \frac{p_i}{1-p_{n+1}} x_i} = \frac{1}{E[Y]} \le E\left[\frac{1}{Y}\right] = \sum_{i=1}^n \frac{p_i}{1-p_{n+1}} \cdot \frac{1}{x_i}. \quad (*)$$

We wish to prove that

$$\frac{1}{\sum_{i=1}^{n+1} p_i x_i} \le \sum_{i=1}^{n+1} p_i \cdot \frac{1}{x_i}.$$

To this end:

$$\begin{aligned}
\frac{1}{\sum_{i=1}^{n+1} p_i x_i} &= \frac{1}{p_{n+1}x_{n+1} + \sum_{i=1}^n p_i x_i} \\
&= \frac{1}{p_{n+1}x_{n+1} + (1-p_{n+1})\sum_{i=1}^n \frac{p_i}{1-p_{n+1}} x_i} \\
&\le \frac{p_{n+1}}{x_{n+1}} + (1-p_{n+1}) \frac{1}{\sum_{i=1}^n \frac{p_i}{1-p_{n+1}} x_i} && \text{by the given inequality} \\
&\le \frac{p_{n+1}}{x_{n+1}} + (1-p_{n+1}) \sum_{i=1}^n \frac{p_i}{1-p_{n+1}} \cdot \frac{1}{x_i} && \text{by } (*) \\
&= \sum_{i=1}^{n+1} p_i \cdot \frac{1}{x_i},
\end{aligned}$$

as desired. ■

(b) [3 pts] Let R, T be positive independent random variables with finitely many outcomes each. Prove that $E[\frac{R}{T}] \ge \frac{E[R]}{E[T]}$.

Solution.

$$\begin{aligned}
E\left[\frac{R}{T}\right] &= E\left[R \cdot \frac{1}{T}\right] \\
&= E[R] \cdot E\left[\frac{1}{T}\right] \\
&= E[R] \cdot \frac{1}{E[T]},
\end{aligned}$$

applying the result of part (b). ■

Problem 6. [8 points] We roll a fair die until we have rolled all 6 numbers. The rolls are independent. What is the expected number of rolls until this happens?

Solution. Let T be the number of rolls to see all 6 numbers. Let t_i be the number of further rolls until we see the next new number after having seen $i-1$ numbers already, so that

$$T = t_1 + t_2 + t_3 + t_4 + t_5 + t_6.$$

Using the linearity of expectation, we can compute $E[T]$ from the expectations $E[t_i]$.

For t_i, we will roll until we see one of the $7-i$ possible remaining numbers. There are 6 faces on the die, so the probability of getting a new number on each roll is $\frac{7-i}{6}$. This lets us compute $E[t_i] = \frac{6}{7-i}$ through a geometric series. Taking the sum,

$$E[T] = E[t_1] + E[t_2] + E[t_3] + E[t_4] + E[t_5] + E[t_6] = 14.7.$$

■

Problem 7. [10 points] We are given a random vector of n distinct numbers. We then determine the maximum of these numbers using the following procedure:

Pick the first number. Call it the *current maximum.* Go through the rest of the vector (in order) and each time we come across a number (call it x) that exceeds our current maximum, we update the current maximum with x.

What is the expected number of times we update the current maximum?

(*Hint: L*)et X_i be the indicator variable for the event that the ith element in the vector is larger than all the previous elements.

Solution. Let's fix the n numbers we are given. We can assume that we are given a random permutation of the n numbers. For $i \in [1, n]$, let X_i be the indicator variable for the event that the ith element in the vector is larger than all the previous elements.

Note that the number of times we update the current maximum is precisely $X_1 + \cdots + X_n$. Since expectation is a linear operator, we can compute $\mathrm{E}\left[X_1 + \cdots + X_n\right]$ by finding $\mathrm{E}\left[X_i\right]$ for each i and summing them up.

Since X_i is an indicator, we only have to find $\Pr\left(X_i = 1\right)$. In a random permutation, $X_i = 1$ happens with probability $1/i$. Why? If you take i distinct numbers and randomly permute them, the probability that the largest one occupies the last (or any given) position is $1/i$.

Thus,

$$\begin{aligned}
\mathrm{E}\left[X\right] &= \sum_i \Pr\left\{X_i = 1\right\} \\
&= \sum_{i=1}^{n} \frac{1}{i} \\
&= H_n \approx \ln n,
\end{aligned}$$

where H_n is the nth Harmonic number. ■

Problem 8. [14 points] Suppose we are trying to estimate some physical parameter p. When we run our experiments and process the results, we obtain an estimator of p, call it p_e. But if our experiments are probabilistic, then p_e itself is a random variable. We call the random variable p_e an *unbiased* estimator if $\mathrm{E}[p_e] = p$.

For example, say we are trying to estimate the height, h, of Green Hall. However, each of our measurements has some noise that is, say, Gaussian with zero mean. So each measurement can be viewed as a sample from a random variable X. The expected value of each measurement is thus $\mathrm{E}[X] = h$, since the probabilistic noise has zero mean. Then, given n independent trials, $x_1, \ldots, x_n$, an unbiased estimator for the height of Green Hall would be

$$h_e = \frac{x_1 + \cdots + x_n}{n},$$

since

$$\mathrm{E}[h_e] = \mathrm{E}\left[\frac{x_1 + \cdots + x_n}{n}\right] = \frac{\mathrm{E}[x_1] + \cdots + \mathrm{E}[x_n]}{n} = \mathrm{E}[x_1] = h.$$

Now say we take n independent observations of a random variable Y. Let the true (but unknown) variance of Y be $\mathrm{Var}[Y] = \sigma^2$. Then we can define the following estimator σ_e^2 for $\mathrm{Var}[Y]$ using the data from our observations:

$$\sigma_e^2 = \frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n} - \left(\frac{y_1 + y_2 + \cdots + y_n}{n}\right)^2.$$

Is this an unbiased estimator of the variance? In other words, is $\mathrm{E}[\sigma_e^2] = \sigma^2$? If not, can you suggest how to modify this estimator to make it unbiased?

Solution. Let $\sigma^2 = \mathrm{Var}[X], \mu = \mathrm{E}[X]$. Then our estimator σ_e^2 is given by

$$\begin{aligned}
\sigma_e^2 &= \frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2 \\
\mathrm{E}\left[\sigma_e^2\right] &= \mathrm{E}\left[\frac{\sum y_i^2}{n} - \left(\frac{\sum y_i}{n}\right)^2\right] \\
&= \frac{\sum \mathrm{E}[y_i^2]}{n} - \frac{\mathrm{E}[(\sum y_i)^2]}{n^2} \\
&= \frac{\sum(\sigma^2 + \mu^2)}{n} - \frac{\mathrm{Var}[\sum y_i] + \mathrm{E}^2[\sum y_i]}{n^2} \\
&= \frac{n(\sigma^2 + \mu^2)}{n} - \frac{n\sigma^2 + n^2\mu^2}{n^2} \\
&= \sigma^2\left(1 - \frac{1}{n}\right)
\end{aligned}$$

So this gives a biased estimator, but we can make it unbiased simply by multiplying by $\frac{n}{n-1}$.

$$\mathrm{E}\left[\frac{n\sigma_e^2}{n-1}\right] = \sigma^2$$

∎