Midterm Solution Sketches

- Don't Panic.
- The midterm contains six problems (and one just for fun). You have 100 minutes to earn 100 points.
- The midterm contains 20 pages, including this one and 4 pages of scratch paper.
- The midterm is closed book. You may bring one double-sided sheet of A4 paper to the midterm. (You may not bring any magnification equipment!) You may not use a calculator, your mobile phone, or any other electronic device.
- Write your solutions in the space provided. If you need more space, please use the scratch paper at the end of the midterm. Do not put part of the answer to one problem on a page for another problem.
- Read through the problems before starting. Do not spend too much time on any one problem.
- Show your work. Partial credit will be given. You will be graded not only on the correctness of your answer, but also on the clarity with which you express it. Be neat.
- Draw pictures and give examples.
- Good luck!

Problem #	**Name**	**Possible Points**	**Achieved Points**
1	A Few Random Variables	10	
2	A Simple Game of Chance	10	
3	In the Dungeon	10	
4	Average Case Insertion Sort	20	
5	Scheduling Exams	15	
6	Random Graphs	35	
Total:		100	

Student Number: ______________________________

Problem 1. A few random variables [10 points]

Let $x_1, x_2, \ldots, x_n$ be a collection of n independent indicator random variables where each $x_j = 1$ with probability p and $x_j = 0$ with probability $(1-p)$. Let $X = \sum_{i=1}^{n} x_i$.

Let S be a subset of k of the random variables, and let E_S be the event that:

$$\forall x_j \in S,\ x_j = 1 \ .$$

For example, if $S = \{x_1, x_2\}$, then $k = 2$ and E_S is the event that the first two random variables are 1.

For each part below, give an exact answer (not an approximation) as a function of n, p, and k.

Problem 1.a. What is $\mathrm{E}[X]$? np

Problem 1.b. What is variance $\mathrm{Var}[X]$? $np(1-p)$

Problem 1.c. What is $\Pr[X = 0]$? $(1-p)^n$

Problem 1.d. For an arbitrary $k > 0$, what is $\Pr[X = k]$? $\binom{n}{k} p^k (1-p)^{n-k}$

Problem 1.e. Given some S and k, what is $\mathrm{E}[X | E_S]$? $(n-k)p + k$

Problem 2. A simple game of chance [10 points]

Alice has a 5-sided die. (Weird, I know.) When you roll the die, each number from 1 to 5 comes up with equal probability. Alice also has a coin. On one side of the coin is the number 2 and on the other side of the coin is the number 4.

5-sided die

coin with 2 and 4 on each side

Problem 2.a. (*Expected Value.*)

What is the expected value of rolling the die? 3

What is the expected value of flipping the coin? 3

Problem 2.b. (*Variance.*)

What is the variance of rolling the die? 4

What is the variance of flipping the coin? 1

Problem 2.c. Alice proposes the following fun game:

- You roll the die and get a number from 1 to 5.
- You flip the coin and get either a 2 or a 4.
- You multiply the results together.

For example, if the die rolls a 5 and the coin flips a 4, then the result is 20.

Alice reasons that we can find the expected value of the game using your answers from the previous part:

> If d is the expected value of the die and c is the expected value of the coin, then $d \cdot c$ is the expected value of the game.

Is Alice's reasoning right or wrong? (Circle your answer.)

RIGHT **WRONG**

Explain your answer: if Alice is correct, explain why; if Alice is wrong, give the correct expected value of the game.

Solution: Alice is right. Since the events are independent, we can conclude that $\mathrm{E}[DC] = \mathrm{E}[D]\,\mathrm{E}[C] = dc$.

Problem 2.d. Alice also reasons that we can find the variance of the game using your answers from the previous part:

> If v_1 is the variance of the die and v_2 is the variance of the coin, then $v_1 \cdot v_2$ is the variance of the game.

Is Alice's reasoning right or wrong? (Circle your answer.)

RIGHT **WRONG**

Explain your answer: if Alice is correct, explain; if Alice is wrong, give the correct variance.

Solution: Alice is wrong. The variance of two events cannot be multiplied together. Assume X and Y are independent events, then:

$$\begin{aligned} \mathrm{Var}[XY] &= \mathrm{E}\left[X^2Y^2\right] - \mathrm{E}[XY]^2 \\ &= \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right] - \mathrm{E}[X]^2\,\mathrm{E}[Y]^2 \end{aligned}$$

It is not true that this is always equal to $\mathrm{Var}[X]\,\mathrm{Var}[Y]$, since $\mathrm{E}[X^2] \neq \mathrm{E}[X]^2$. In this case, we can directly calculate:

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= (1+4+9+16+25)/5 = 11 \\ \mathrm{E}\left[Y^2\right] &= (4+16)/2 = 10 \end{aligned}$$

$$\mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right] = 110$$

$$\mathrm{E}[X]^2\,\mathrm{E}\left[Y^2\right] = 9 \cdot 9 = 81$$

Therefore, the expected variance of the game is $110 - 81 = 29$.

Problem 3. In the dungeon [10 points]

Redmore the Red, an adventurer of yore, is arrested by the king and thrown in the dungeon. Also in the dungeon are two other adventurers, Alice and Bob.[1] The prison guard informs them that one of them has been selected, uniformly at random, for execution.

Redmore is naturally quite afraid, and he asks the guard which of them is to be executed. The guard responds that he is not allowed to tell Redmore his fate, but that he would provide the following information:

- If Redmore *was not* selected for execution, then the guard will tell Redmore accurately whether Alice or Bob is to be executed.
- If Redmore *was* selected for execution, then the guard will randomly flip a coin and with probability 1/2 indicate Alice, and with probability 1/2 indicate Bob.

Redmore agrees to this plan. (He did not have much choice, did he?) The guard then tells Redmore that Alice is going to be executed.

Given the guard's information:

- What is the probability that Alice is executed? [2/3]
- What is the probability that Bob is executed? [0]
- What is the probability that Redmore is executed? [1/3]

Explain your answer:

Solution: Let A, B, and R be the events that Alice, Bob, and Redmore are selected for execution, respectively. We know that $\Pr[A] = \Pr[B] = \Pr[R] = 1/3$.
Let G_A be the event that the guard says Alice. We calculate this probability as follows:

$$\begin{aligned}\Pr[G_A] &= \Pr[G_A|R]\Pr[R] + \Pr[G_A|A]\Pr[A] + \Pr[G_A|B]\Pr[B] \\ &= (1/2)(1/3) + (1)(1/3) + (0)(1/3) \\ &= 1/2\end{aligned}$$

We can now calculate $\Pr[R|G_A] = \Pr[R \cap G_A]/\Pr[G_A]$. We know that $\Pr[R \cap G_A] = 1/6$, and hence $\Pr[R|G_A] = 1/3$.
We can similarly calculate that $\Pr[A|G_A] = \Pr[A \cap G_A]/\Pr[G_A] = 2/3$, and $\Pr[B|G_A] = 0$.

[1] Alice was arrested for cheating in a weird game involving a five-sided die.

Problem 4. Average-Case Analysis of InsertionSort [20 points]

One of the first methods that we teach for sorting is *InsertionSort*. The algorithm proceeds through the array, sorting each element into place, one at a time, by comparing to all the preceding items in the list. The details are unimportant for this problem, but the pseudocode is as follows:

```
InsertionSort(Array A, integer n)
    for i=1 to (n-1) do
        item = A[i];
        int slot = i;
        while (slot > 0) and (A[slot] > item) do
             A[slot] = A[slot-1];
             slot = slot-1;
        A[slot] = item;
}
```

The key property that is important is that the running time of InsertionSort depends on the number of *inversions* in the permutation being sorted. Given a sequence of integers $\{a_1, a_2, \ldots, a_n\}$, an **inversion** is a mis-ordered pair (a_j, a_k) where $j < k$ but $a_j > a_k$. For example, in the sequence:

$$S_1 = \{1 \;\; 2 \;\; 5 \;\; 3 \;\; 10\}$$

there is one inversion: the pair $(5, 3)$. As another example, consider the sequence:

$$S_2 = \{1 \;\; 2 \;\; 10 \;\; 5 \;\; 3\}$$

This sequence contains three inversions: $(10, 5)$, $(10, 3)$, and $(5, 3)$. The following theorem relates the running time of InsertionSort to the number of inversions in the input:

Theorem 1 *If a permutation S of length n contains I inversions, then* `InsertionSort(S,n)` *runs in time $\Theta(n + I)$.*

Continued on the next page.

Analyze the *average-case* performance of InsertionSort. Let S be a permutation of the integers $[1, \ldots, n]$ chosen uniformly at random from the set of all permutations. Show that the expected running time of InsertionSort on S is $\Theta(n^2)$.

Solution to Problem 5:

Solution: Let $S = \{a_1, a_2, \ldots, a_n\}$ be the random input sequence. Fix two elements of the sequence i and j. Without loss of generality, assume $i < j$. Since the input sequence is a random permutation, $\Pr[a_i > a_j] = 1/2$. That is, elements i and j create an inversion with probability 1/2.

For every pair (i, j) where $i < j$, let $X_{i,j}$ be the indicator random variable that equals 1 if $a_i > a_j$ and 0 otherwise. We know that $\Pr[X_{i,j}] = 1/2$, and hence $\mathrm{E}[X_{i,j}] = 1/2$.

Let X be the total number of inversions in the random input sequence. Notice that $X = \sum_{i<j} X_{i,j}$. Thus:

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}\left[\sum_{i,j} X_{i,j}\right] \\
&= \sum_{i<j} \mathrm{E}[X_{i,j}] \\
&= \sum_{i<j} \frac{1}{2} \\
&= \frac{n(n-1)}{2} \cdot \frac{1}{2} \\
&= \frac{n(n-1)}{4}
\end{aligned}$$

Thus, the expected number of inversions is $\Theta(n^2)$, and hence the expected running time of InsertionSort is $\Theta(n + n^2) = \Omega(n^2)$.

Problem 5. Scheduling exams [15 points]

Imagine a university with n students and m modules. Each module has a final exam. And each student takes at least k different modules. (Some may take more than k modules!) You may assume that $k \geq c \log n$ for some constant c. (Do not worry if c is a large integer.)

The exam period consists of two weeks: Week A and Week B. Each exam is randomly assigned to Week A with probability 1/2 or Week B with probability 1/2.

Your goal is to prove that, with high probability, no student has too many exams in either week: if a student is taking x exams, then no more than 3/4 of her exams should be in Week A and no more than 3/4 of her exams should be in Week B.

Example: Perhaps Alice is taking {CS1001, CS2001, CS3001, and CS4001}. Perhaps Bob is taking {CS2001, CS2002, CS2003, CS2004}. Perhaps the random assignment assigns to Week A the following exams: {CS1001, CS2001, CS2002, CS2003}; it assigns to Week B the following exams: {CS2004, CS3001, CS4001}.

In this case, both students are happy since they have $\leq 3/4$ of their exams in each week: Alice has two exams in Week A and two exams in Week B; Bob has three exams in Week A and one exam in Week B.

Problem 5.a. Give a one sentence overview of your proof strategy.

Solution: We fill fix a student and a week and use a Chernoff Bound to show that not too many exams are assigned to that week; we will then take a union bound over all students and all weeks.

Problem 5.b. Define the random variables that you will use.

Solution: For a fixed student (say, Alice) and a fixed week (say, Week A), if the student is taking ℓ exams then we will define ℓ indicator random variables $x_1, x_2, \ldots, x_\ell$ as follows: $x_j = 1$ if the student's j^{th} exam is assigned to this week; $x_j = 0$ otherwise.

Continued on the next page.

Problem 5.c. What probabilistic tools / tail bounds / etc. will you use to prove this? Give the complete definition of the bound (not just its name).

Solution: We will use the following Chernoff Bound: Given a set of random variables $x_1, \ldots, x_\ell$ where $X = \sum_j x_j$ and $\mu = \mathrm{E}[X]$, and given $\delta < 1$:

$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}$$

We will also use a union bound: Given a set of events $E_1, E_2, \ldots, E_n$ where each event occurs with probability at least p, then the probability that *any* such event occurs is at most np.

Problem 5.d. Recall that we assumed each student is taking at least $k \geq c \log n$ modules, for some constant c. What value of c will you require in your proof? (You can choose any constant. Do not worry if it is unrealistically large.)

Choose $c =$ 48 **Solution:** Choose $c = 48$.

Problem 5.e. Give the complete proof that with probability at least $(1 - 2/n)$, every student has no more than $3/4$ of their exams in either week.

Solution: As specified above, we have fixed a student and a week, and define the indicator random variables as specified. Assume the student is taking $\ell \geq k \geq 48\log(n)$ exams. We know that $\mathrm{E}[x_j] = \Pr[x_j = 1] = 1/2$. Define $X = \sum_j x_j$, and $\mu = \mathrm{E}[X] = \ell/2$.

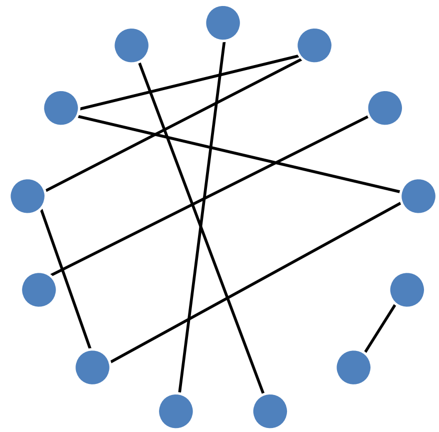
By a Chernoff Bound (see above) we know that:

$$\begin{aligned}
\Pr[X \geq (3/4)\ell] &= \Pr[X \geq (1+1/2)\mu] \\
&\leq e^{-\ell/2(1/4)(1/3)} \\
&\leq e^{-k/24} \\
&\leq e^{-48\log n/24} \\
&\leq 1/n^2
\end{aligned}$$

That is, for this student and this week, the probability that there are more than 3/4 of the student's exams in this week is at most $1/n^2$.

In total, there are n students and 2 weeks, so there are $2n$ possible events $E_{1A}, E_{1B}, E_{2A}, E_{2B}, \ldots$ where event E_{iA} indicates that student i has too many exams in week A, etc. We have just proved that each such event occurs with probability at most $1/n^2$. Thus, by a union bound, we know that *any* of these events occurs with probability at most $2n/n^2 \leq 2/n$. This completes the proof.

Problem 6. Random Graphs [30 points]



Imagine we have a random graph $G(n,p)$ that we construct as follows:

- Begin with n nodes and no edges. Let V be the set of n nodes. Let edges $E = \emptyset$.
- For each pair of nodes $u, v \in V$, add (undirected) edge (u,v) to E with probability p.

Problem 6.a. Assume $p = \frac{18 \log n}{n}$. Prove that with probability at least $1 - 1/n$, every node in graph $G(n,p)$ has degree at most $O(\log n)$.

Solution: We can solve this using a Chernoff Bound (followed by a union bound). Fix a specific node u. For each other node in the graph, there is a probability p of adding an edge. Let x_j be the indicator random variable where $x_j = 1$ if u has an edge to node j. Let $X = \sum x_j$ be the degree of u. The expected degree of u is equal to $\mathrm{E}\left[\sum x_j\right] = p(n-1) \leq pn \leq 18 \log n$. Also notice that $p(n-1) \geq 18\log(n)/2 \geq 9 \log n$. By a Chernoff Bound, we conclude that:

$$\begin{aligned} \Pr\left[X \geq 36 \log n\right] &\leq \Pr\left[X \geq (1+1)\mathrm{E}\left[X\right]\right] \\ &\leq e^{-p(n-1)/3} \\ &\leq e^{-9\log n/3} \\ &\leq 1/n^2 \end{aligned}$$

By a union bound, we conclude that every node in n has a degree at most $36 \log n$ with probability at least $1 - 1/n$.

Problem 6.b. Assume $p = \frac{1}{3n}$. Prove that with probability at least 1/2, graph $G(n, p)$ is *not connected*. (You may assume that n is sufficiently large, e.g., $n > 10$.)

Hint: What is the expected number of edges?

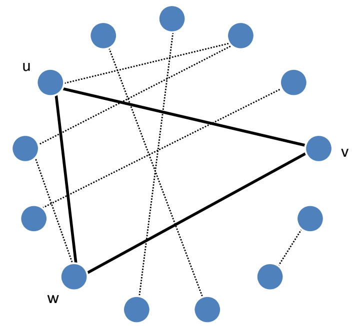
Solution: First, we observe that the expected number of edges is $p\binom{n}{2} \leq pn^2 \leq n/3$, by linearity of expectation. Thus, by Markov's Inequality:

$$\Pr\left[|E| \geq n-1\right] \leq \mathrm{E}\left[|E|\right]/(n-1) \leq \frac{n}{3(n-1)} \leq \frac{n}{2n} \leq 1/2$$

(Notice that here we assumed that $3(n-1) \geq 2n$, which holds as long as $n \geq 3$. Of course, if $n = 2$, then the claim follows trivially!) Thus we have shown that with probability at least 1/2, the graph $G(n, 1/3n)$ is disconnected. Notice, of course, that this is also true if $p = 1/n$ since the total number of possible edges is actually $n(n-1)/2$.

Problem 6.c. How many triangles are there in $G(n, p)$?

A triangle is a set of nodes $u, v, w \in V$ where all three nodes are connected. That is, we say that $T = (u, v, w)$ is a triangle in graph $G = (V, E)$ if: $(u, v) \in E, (v, w) \in E, (u, w \in E)$.



Consider the random graph $G(n, p)$ where $p = \frac{1}{2n}$. Show that with probability at least 1/2, graph $G(n, p)$ has no triangles.

Solution: For a given triangle $T = (u, v, w)$, graph G contains triangle T with probability p^3. There are $\binom{n}{3} \leq n^3$ triangles. Thus the total expected number of triangles is at most $n^3p^3 \leq 1/8$. Let C be the number of triangles in $G(n, p)$. By Markov's Inequality:

$$\Pr[C \geq 1] \leq \mathrm{E}[C] \leq 1/8 \ .$$

So with probability at least 7/8, graph $G(n, 1/2n)$ has no triangles.

Problem 6.d. Now we want to show the opposite claim: that if $p = \frac{\alpha}{n}$, for some constant $\alpha > 1$, then graph $G(n, p)$ has at least one triangle with constant probability. We will do this in two steps: first, we will bound the variance on the number of triangles, and then we will do the actual probability calculation. Assume here that α is some constant (e.g., $\alpha \geq 10$), and that n is sufficiently large (e.g., $n > 100\alpha^{10}$).

Let C be the random variable representing the number of triangles in $G(n, p)$. Assume that we have already proven that the variance $\mathrm{Var}[C] \leq 2\alpha^3$. Use Chebychev's Inequality to show that with probability at least 1/2, graph $G(n, p)$ has at least one triangle.

Solution: First, we observe that $\mathrm{E}[C] \geq p^3 \binom{n}{3} \geq \alpha^3(1 - 2/n)^3 > \alpha^3/8$ for $n \geq 4$.

$$\begin{aligned} \Pr[|C - \mathrm{E}[C]| \geq \mathrm{E}[C]] &\leq \frac{\mathrm{Var}[C]}{\mathrm{E}[C]^2} \\ &\leq \frac{2\alpha^3}{\alpha^6/64} \\ &\leq \frac{128}{\alpha^3} \\ &\leq \frac{1}{2} \end{aligned}$$

From this we conclude that with probability at least 1/2, $|C - \mathrm{E}[C]| < \mathrm{E}[C]$ and hence $C > 0$. That is, with probability at least 1/2, there is at least 1 triangle in $G(n, p)$.

Problem 6.e. Prove that variance $\mathrm{Var}[C] \le 2\alpha^3$.

Hint 1: For a triangle T, let x_T be an indicator random variable where $X_T = 1$ if triangle T is in the graph $G(n,p)$ and $x_T = 0$ otherwise. Let $C = \sum_T x_T$. The variance of C is:

$$\begin{aligned} \mathrm{Var}[C] &= \sum_{T_1}\sum_{T_2} \mathrm{CoVar}[T_1, T_2] \\ \mathrm{CoVar}[T_1, T_2] &= \mathrm{E}[x_{T_1} x_{T_2}] - \mathrm{E}[x_{T_1}]\,\mathrm{E}[x_{T_2}] \end{aligned}$$

Hint 2: You may want to divide this summation into three components: the set of triangle pairs (T_1, T_2) that share: (i) zero edges, (ii) one edge, and (iii) three edges.

Hint 3: In some cases, you also may find it useful to observe that $\mathrm{CoVar}[T_1, T_2] \le \mathrm{E}[x_{T_1} x_{T_2}]$.

Solution: We will consider the three cases separately. First, consider the case where T_1 and T_2 are disjoint. In this case, x_{T_1} and x_{T_2} are independent. Thus $\mathrm{E}[x_{T_1} x_{T_2}] = \mathrm{E}[x_{T_1}]\,\mathrm{E}[x_{T_2}]$, and so the covariance of these triangle pairs is 0.

Next, consider a triangle pair in C_1, i.e., the case where T_1 and T_2 share one edge. These two triangles contain 5 edges in total, and so the probability that both occur is p^5, i.e., $\mathrm{CoVar}[T_1, T_2] \le \mathrm{E}[x_{T_1} x_{T_2}] \le p^5$. There are at most n^2 edges that might be shared by T_1 and T_2, and there are at most n^2 ways to choose two more nodes to combine with the edge to get two triangles. Hence there are at most n^4 triangle pairs in the set C_1. Hence the sum of expectations for these triangle pairs in C_1 is at most $n^4 p^5 \le \alpha^5/n$.

Next, consider a triangle pair in C_3, i.e., the case where T_1 and T_2 share all three edges. (That is, the two triangles are identical.) In this case, the probability of all three edges being added to the graph G is p^3, and there are at most n^3 such "pairs" of triangles T_1 and T_2 in C_3. Thus we conclude that the sum of the covariance of these triangle pairs in C_3 is $p^3 n^3 \le \alpha^3$.

Thus we conclude that $\mathrm{Var}[C] \le \alpha^5/n + \alpha^3 \le 2\alpha^3$ for $n \ge \alpha^2$.

Note: You will observe that we have shown something interesting here! The probability $1/n$ is a critical threshold for the random graph $G(n,p)$. If $p > 1/n$, then we are likely to have one triangle; if $p < 1/n$ we are likely to have no triangles. And a triangle is really just a cycle of length 3. What if we considered a cycle of length k? Does graph $G(n,p)$ contain a cycle of length k? Would it be surprising if $1/n$ were a critical threshold for this question as well?

Scratch Paper

Scratch Paper

Scratch Paper

Scratch Paper