MATH 341: PROBABILITY: FALL 2009
COMMENTS ON HW PROBLEMS

STEVEN J. MILLER (SJM1@WILLIAMS.EDU)

ABSTRACT. A key part of any math course is doing the homework. This ranges from reading the material in the book so that you can do the problems to thinking about the problem statement, how you might go about solving it, and why some approaches work and others don't. Another important part, which is often forgotten, is how the problem fits into math. Is this a cookbook problem with made up numbers and functions to test whether or not you've mastered the basic material, or does it have important applications throughout math and industry? Below I'll try and provide some comments to place the problems and their solutions in context.

1. HW #1

The first assignment was: Due Thursday, 9/17 (but as this is the first assignment, no late penalty if you put it in my mailbox by 10am on Friday the 18th): Section 1.3: #2, #3, #5; Combinatorics: (1) There are 2n people who enter as n pairs of two. The people are then randomly matched in pairs. What is the probability everyone is matched with their initial partner? There are two ways to interpret this problem; either is fine so long as you state which interpretation. In one interpretation, say there are n people from Williams and n from Amherst, matched in n pairs with each pair having someone from Williams and someone from Amherst. In the new matching, you must match someone from Williams with someone from Amherst. In the other interpretation, anyone can be matched with anyone. You may solve either problem, just clearly state which one you are doing (not surprisingly, the answers differ). (2) Consider n people ordered 1, 2, ..., n. We randomly assign another ordering to these people – what is the probability at least one person is assigned the same number twice? Section 1.4: #2, #4. Section 1.8: #2, #4, #6, #12.

Section 1.3: Problem 2: This problem on Murphy's law is quite important, and will be used later for the elementary analysis of the symmetric random walk (also known as the Gambler's ruin). If we consider a sequence of k tosses, then the probability it is observed when we toss a fair coin k times is just $p = 1/2^k$, or the probability it does not happen is $1 - p = 1 - 1/2^k$. If we toss a fair coin kN times, then the probability it does not happen in one of these blocks is $(1-p)^N$, which tends to 0 as $N \to \infty$. Note that the sequence could still occur even if it doesn't occur entirely in one block. For example, say our sequence is TTHT. Imagine we toss the coin 20 times, and get

$$\text{THHT THTH THHT TTTH HHTH.}$$

Date: December 8, 2009.

Note none of the blocks of four have the sequence TTHT, but it does occur in the sequence of 20 (part in the first and part in the second blocks).

Section 1.3: Problem 3: For this problem, the trick is to just enumerate so that you cover all possibilities. This problem is more to test understanding of the material instead of applications for later. I find it is easiest to give each cup a label (so red cup 1, red cup 2). We might as well place the six saucers in order: red saucer 1, red saucer 2, ..., star saucer 2. There are 6! ways to arrange the 6 cups on the saucers (we ARE distinguishing between which red cup is place on a given saucer). To count how many ways to place the cups so that nothing is placed on the same color, there are three possibilities: the two reds are placed on the two whites, the two reds are placed on the two stars, or one red on a white and one red on a star. The problem is completed by counting all configurations like this.

Section 1.3: Problem 5: This problem can be interpreted as saying that if we have a countable collection of events and each event happens with probability 1, then their intersection happens with probability 1. The simplest way to prove this is by induction. If X and Y happen with probability one, then $\mathbb{P}(X \cap Y) = \mathbb{P}(X) + \mathbb{P}(Y) - \mathbb{P}(X \cup Y)$. Note every probability on the right hand side equals 1 (no event can have probability greater than 1, and $X \subset X \cup Y$ so $\mathbb{P}(A \cup Y) = 1$). This implies $\mathbb{P}(X \cap Y) = 1$. Proceed by induction, setting $X = \cup_{r=1}^{n} A_r$ and $Y = A_{n+1}$ to get $\mathbb{P}(\cap_{r=1}^{n+1} A_r) = 1$ for all n. The proof is completed by invoking Lemma 5 on page 7. We could have argued slightly differently above. The key is proving $\mathbb{P}(X \cap Y) = 1$; another approach is to use partitions, and observe $\mathbb{P}(X) = \mathbb{P}(X \cap Y) + \mathbb{P}(X \cap Y^{c})$. As $\mathbb{P}(Y) = 1$, $\mathbb{P}(Y^{c}) = 0$ and thus $\mathbb{P}(X \cap Y^{c}) = 0$ (as $X \cap Y^{c} \subset Y^{c}$). Thus $\mathbb{P}(X) = \mathbb{P}(X \cap Y)$, and as $\mathbb{P}(X) = 1$ we finally deduce $\mathbb{P}(X \cap Y) = 1$. ***Note how important in this problem the $n = 2$ case is in the inductive proof. Frequently in induction proofs we just need to use the result with n to prove $n + 1$; however, a sizable number of times the general proof basically just reduces to understanding the $n = 2$ case.***

Combinatorics Problem (1): If anyone can be matched with anyone, there are $(2n - 1)!!$ ways to do this, where the double factorial means we take the product of every other term ($6!! = 6 \cdot 4 \cdot 2$ and $5!! = 5 \cdot 3 \cdot 1$). One way to see this is to note this is just

$$\binom{2n}{2}\binom{2n-2}{2}\cdots\binom{4}{2}\binom{2}{2} \cdot \frac{1}{n!};$$

we divide by $n!$ as we have attached labels to each pair of people, and there aren't supposed to be labels. We could also proceed by induction. The first person must be matched with someone; there are $2n - 1$ ways to do this. We now pair off the remaining $2n - 2$ people, which by induction happens $(2n - 3)!!$ ways, so there are $(2n - 1) \cdot (2n - 3)!! = (2n - 1)!!$ ways. If you must be matched with someone from the opposite side, there are only $n!$ ways.

Combinatorics Problem (2): We solve this by inclusion-exclusion. Let A_i be the event that i is in the i^{th} place, A_{ij} be that i and j are in their respective places (with $i \neq j$), and so on. Note $\mathbb{P}(A_{134}) = \mathbb{P}(A_{459})$ and so on. Then the probability that at least one person is in the right spot is

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k=1}^{n} A_i\right) &= \sum_{i\leq n}\mathbb{P}(A_i) - \sum_{i<j\leq n}\mathbb{P}(A_{ij} + \sum_{i<j<k\leq n}\mathbb{P}_(A_{ijk} - \cdots \\ &= \binom{n}{1}\mathbb{P}(A_1) - \binom{n}{2}\mathbb{P}(A_{12}) + \binom{n}{3}\mathbb{P}(A_{123}) - \cdots \\ &= \frac{n}{1!}\frac{1}{n} - \frac{n(n-1)}{2!}\frac{n(n-1)}{+}\frac{n(n-1)(n-2)}{3!}\frac{1}{n(n-1)(n-2)} - \cdots \\ &= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!}.\end{aligned}$$

As $n \to \infty$, this tends to $1 - 1/e$; this follows from algebra applied to the Taylor series expansion of e^x with $x = -1$.

Section 1.4: Problem 2: This is another good problem to use induction, where again the key step is when $n = 2$. By the definition, we have $\mathbb{P}(X \cap Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$. The base case is taking $X = A_2$ and $Y = A_1$. In general, we set $X = A_{n+1} \cap \cdots \cap A_2$ and $Y = A_1$, and the result follows by induction.

Section 1.4: Problem 4: No one asked me about this, so assuming all is good.

Section 1.8: Problem 2: To have exactly two kings and one ace in 13 cards means we choose 2 of 4 kings, 1 of 4 aces, and then 10 of 44 non-king and non-aces. Thus the number of possible hands is $\binom{4}{2}\binom{4}{1}\binom{44}{10}$; as there are $\binom{52}{13}$ ways to choose 13 cards for the hand, the probability is just the ratio. As no one asked about this problem, I'll assume the second part is fine. ***For problems like this, it is very easy to double count. The danger is getting a third king or a second ace. I find it is easiest to break it up like this, where we first go through the kings, then the aces, then the remaining. note the numbers in the n choose r's up top add to 52 and the bottom adds to 13.***

Section 1.8: Problem 4: For (a), $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$ (there are 2^8 possibilities; it is important to enumerate in such as way that none are missed). There are many choices for the σ-field. The simplest is to take $\mathcal{F} = \{\varphi, \Omega\}$; while this satisfies all the requirements of a σ-field, it is a very poor choice. It allows us to only talk about probabilities of nothing happening or something happening. The larger the σ-field, the better. If Ω is finite or countable, we can and should take the σ-field to be 2^Ω, the set of all subsets of Ω. As Ω has 8 elements, here there would be $2^8 = 256$ elements in the σ-field. Some of these are $\varphi, \Omega, \{HHH, TTT\}, \{HHT, THT, TTT\}$ and so on. Finally we must define a measure. If we can define a probability on each element $\omega \in \Omega$ then we can define the probability of an A in the σ-field by $\mathbb{P}(A) = \sum_{\omega\in A}\mathbb{P}(\omega)$. This is very important, as we would hate to have to define the probability of each of the 256 possible subsets of Ω directly; defining the probabilities of the singletons of Ω induces the probabilities elsewhere. As we are told the coin is biased, let p be the probability of a head, and then $\mathbb{P}(\omega) = p^{\#H(\omega)}(1-p)^{1-\#H(\omega)}$, where $\#H(\omega)$ is the number of heads in ω (thus

$\mathbb{P}(HTH) = p^2(1-p)$). If we had an uncountable Ω, we couldn't do this. For us, if we have $[0,1]$, $[0,1]^n$, $\mathbb{R}$ or $\mathbb{R}^n$ (or anything like that), we take for the σ-field the set of subsets of Ω generated by open intervals.

Section 1.8: Problem 6: As no one asked about this, I'll assume all is good.

Section 1.8: Problem 12: Not surprisingly, this is another induction problem where the key observation is using the $n = 2$ claim. We have $\mathbb{P}(X \cap Y) = \mathbb{P}(X) + \mathbb{P}(Y) - \mathbb{P}(X \cup Y)$. Now take $X = A_1 \cap \cdots \cap A_n$ and $Y = A_{n+1}$. The algebra becomes a bit tedious, but we have

$$\mathbb{P}\left(\bigcap_{k=1}^{n+1} A_k\right) \ = \ \mathbb{P}\left(\bigcap_{k=1}^{n} A_k\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left((A_1 \cap \cdots \cap A_n) \cup A_{n+1}\right).$$

In the expansion above, the difficult part is the last piece. It is a mix of intersections and unions, and our desired formula only has unions on the right. The solutions is to note

$$(A_1 \cap \cdots \cap A_n) \cup A_{n+1} \ = \ (A_1 \cup A_{n+1}) \cap \cdots \cap (A_n \cup A_{n+1});$$

to see this, argue as follows: either $x \in A_{n+1}$ or $x \in A_k$ for all $k \le n$. Now we have a probability of the union of n sets, and can expand. ***In induction after induction, we see the advantage of grouping terms and using the results from the*** $n = 2$ ***case.***

---

2. HW #2

Homework: Due Thursday 9/24 (though you may place in my mailbox anytime up till 10am on Friday 9/25): Section 1.5: #1, #2, #4 (also determine if it is true if p is not prime), #8. Section 1.7: #1, #3 (hint: you can solve this without using difference equations!), #4. Section 1.8: #28 (also determine if we must have 10% colored, or if we can do more, and generalize to 4-dimensions if possible). Section 2.1: #2, #4, #5c. Section 2.3: #3 (very important problem for simulating random variables), #4, #5.

Section 1.5: Problem 1: Lots of ways to do this problem. Easiest is probably to first show that if X and Y are independent then so too are X and Y^{c}. We can now reason and get A^{c} and B^{c} are independent as follows: A, B independent implies A and B^{c} are independent; we then take $X = B^{c}$ and $Y = A$ to get B^{c} and A^{c} are independent. This is a nice trick, marching down like so.

Section 1.5: Problem 2: Consider (assuming $n \ge 3$) the events A_{12}, A_{23} and A_{13}. If the first two happen, then the first and second rolls are the same as well as the the second and third. Thus, the first and third *must* be the same!

Section 1.5: Problem 4: The primality of p is very important. If the events are independent, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Here, $\mathbb{P}(C) = c/p$ where c is the number of elements of C, and thus is an integer between 0 and p (remember the probability of C is just the cardinality of C divided by p). If A and B are independent, then

$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Letting N_C denote the number of elements in C, this means $N_{A\cap B}/p = N_A N_B/p^2$, or $pN_{A\cap B} = N_A N_B$. As p is prime, if p divides a product it must divide one of the two factors, and thus either $p|A$ or $p|B$. Without loss of generality, assume $p|A$; there are only two ways this can happen, namely either A is empty or A is the entire space. What is p is not prime? Let $p = 6$ and set $\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2, 3\}$ and $B = \{3, 4\}$. Then $\mathbb{P}(A) = 3/6$, $\mathbb{P}(B) = 2/6$ and $\mathbb{P}A \cap B = 1/6$. Thus it is essential that p be prime. ***In problems like this, it is a very good idea to ask how important a condition is. Frequently the result is either false if the condition fails, or the proof is much harder.***

Section 1.7: Problem 1: This is a computational problem to make sure you have understood the section. I think the final answer is something like $\frac{1-p^2}{2-p^2}$. Whenever you do a problem, it is worthwhile trying to get a feel for the answer. Is it reasonable? What kind of tests can we do? Well, first off we need to make sure the answer is between 0 and 1, as it is a probability – this is always the case, and thus so far so good. (At least one student showed me a calculation with an error; I was able to easily find the error in one place because the resulting probability exceeded 1.) Is the answer reasonable as p approaches natural limits? If $p \to 1$ then the probability tends to 0; this is as expected, for in this case *all* roads are blocked. What about $p \to 0$ – is 1/2 reasonable? Yes: in this case it is very rare for roads to be blocked, and thus only roads that must be blocked are. If we are told that there is no path from A to C, then it is just as likely that there are two blocked roads from A to B as from B to C. ***In problems such as this, you should always do simple tests like this to see how reasonable an answer is. Is the probability between 0 and 1? What can you say about the answer in extreme cases / limits? Try to tell a story: if*** $p \to 1$ ***then all roads are blocked so....***

Section 1.7: Problem 3: One solution is to use difference equations. The probability we eventually reach N is just k/N, while the probability we reach 0 is $1 - k/N$ (by symmetry); thus the probability neither of these two events happens is $1 - (k/N) - (1 - k/N) = 0$. This solution is somewhat unsatisfying, as it requires us to solve difference equations to get the probability k/N (and the difference equation doesn't even have distinct roots to its characteristic polynomial!). We can find a more elegant solution by using the Murphy's Law problem from Section 1.3 (problem #2). Imagine we flip our fair coin N times and get all tails. Then no matter where we are in $\{1, 2, \dots, N-2, N-1\}$, after N tails we must hit the boundary of 0. As 'eventually' we get N consecutive tails, we must eventually either hit 0 or have been absorbed at N.

Section 1.7: Problem 4: I agree: here's one approach. Let A be the event that we prefer x to y. Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap C) + \mathbb{P}(A \cap C^{\mathrm{c}}) \\ &= \mathbb{P}(A|C)\mathbb{P}(C) + \mathbb{P}(A|C^{\mathrm{c}})\mathbb{P}(C^{\mathrm{c}}) \\ &= 1 \cdot \mathbb{P}(C) + 1 \cdot \mathbb{P}(C^{\mathrm{c}}) \ = \ 1,\end{aligned}$$

where above we used $\mathbb{P}(A|C) = 1$ and $\mathbb{P}(A|C^{\mathrm{c}}) = 1$ as we were told that given C (respectively, C^{c}) we prefer x to y. ***This is a very important problem. When we try***

to prove results, it is often easier to break into cases that cover all possibilities, as then we get to assume additional results in each case. For example, we might have a result is true if the Riemann hypothesis holds (where we use certain consequences of the Riemann hypothesis holding), as well as it is true if the Riemann hypothesis is false (where we use certain consequences of its failure); thus the claim must be true, as either the Riemann hypothesis is true or it is false. This is a powerful way to attack many problems, as in each case we now have a lot more at our disposal. (For those interested, I'm thinking of the proof of Skewes' number; see `http://en.wikipedia.org/wiki/Skewes_number`.

Section 1.8: Problem 28: Each cube has 8 vertices on the surface of the sphere. If I try to be as obstructionist as possible, I will arrange it so that exactly one of the 8 vertices on each cube is blue; this uses as little paint as possible to block you as much as possible. Note no vertex is on two distinct cubes, so without loss of generality I might as well assume that I always paint the vertex in the positive octant blue. I thus need to paint at least 1/8 or 12.5% of the sphere blue to ensure that there is no cube where all vertices are red, but I only have enough paint to do 10%. The generalization to higher dimensions is actually straightforward – all that matters is the number of vertices! In four dimensions we have $2^4 = 16$ vertices, so any number less than 1/16 will suffice to ensure that we can find a hypercube with all vertices painted red. ***This is a red-herring problem. It seems that for higher dimensions we'll need to know things such as the hypervolume and surface area of the spheres, but all we need to know is the number of vertices on the hypercube. Whenever doing a problem, think about the key features of the problem – what really matters, and what might be misleading. We do need to have some of the symmetry of the sphere (it is important coloring one vertex blue cannot block multiple cubes), but we do not need any fine properties of the sphere.***

Section 2.1: Problem 2: $\mathbb{P}(Y \le y)$ is the same as $\mathbb{P}(aX + b \le y)$ or $\mathbb{P}(X \le \frac{y-b}{a}) = F\left(\frac{y-b}{a}\right)$. ***For problems like this, I find it is best to go slow. Start with the definition of the (cumulative) distribution function of*** Y***, and then do some algebra to express this in terms of the (cumulative) distribution function of*** X***.***

Section 2.1: Problem 4: This is a straightforward calculation, so long as $\lambda \in [0, 1]$ (if not the properties break down; for instance, it need not assign non-negative probabilities to intervals). The product is a distribution function (this can be seen after some algebra). ***Similar to a previous problem, we should ask how important is the condition that*** $\lambda \in [0, 1]$***. It turns out that the result is false. If*** $\lambda = 4$***, we have*** $H(x) = 4F(x) - 3G(X)$***. This satisfies the right behavior as*** $x \to \pm\infty$***, but can give negative probabilities. For example, if*** F ***arises from the uniform distribution on*** $[2, 3]$ ***and*** G ***from the uniform on*** $[0, 1]$***, then*** $H(1) = -3$***.***

Section 2.1: Problem 5c: Let $F'(x) = f(x)$. There is no problem with the limits as $x \to \pm\infty$ so long as we remember that $u \log u \to 0$ as u tends to 0 from above. We must show that this function is non-decreasing to complete the proof that it is a distribution function. The simplest way to see this is to take a derivative, which gives

$$f(x) - f(x)\log(1 - F(x)) + (1 - F(x))\frac{-f(x)}{1 - F(x)} \;=\; -f(x)\log(1 - F(x));$$

as $F(x) \in [0, 1]$ we have $\log(1 - F(x)) < 0$ (it is the logarithm of a negative number), and thus the first derivative is positive (so the function is increasing. ***I find the above is a very useful way to prove certain types of claims. Namely, take the first derivative and show it is positive – this suffices to give strictly increasing.***

Section 2.3: Problem 3: This is perhaps one of the most important problems in the entire course! As F is continuous and strictly increasing, it has a continuous inverse F^{-1}. Note $\mathbb{P}(Y \le y) = \mathbb{P}(F^{-1}(X) \le y)$; however, $F^{-1}(X) \le y$ means $X \le F(y)$. Then $\mathbb{P}(Y \le y)$ equals $\mathbb{P}(X \le F(y))$; as $F(y) \in [0, 1]$, from the givens of the problem $\mathbb{P}(X \le F(y)) = F(y)$, which completes the proof. ***Why is this problem so important? One way of interpreting the result is to say that if we can simulate any random variable that is uniformly distributed (or equidistributed) on*** $[0, 1]$***, then we can simulate any random variable whose cumulative distribution function is strictly increasing. Of course, how does one generate a random number uniformly? This is a very hard question. See for instance `http://www.random.org/`.***

Section 2.3: Problem 4: The first part is a computation. The second is false. The easiest example is f is the uniform density on $[0, 1]$ and g the uniform density on $[2, 43]$. Then $f(x)g(x) = 0$ for all x. ***It's often a good idea to play around searching for counterexamples, or seeing what makes examples succeed. Just because*** f ***and*** g ***are non-negative and integrate to 1, nothing implies the same must be true for their product.***

Section 2.3: Problem 5: This is a calculation, and a test of Math 103/104. For (a), note the integral diverges unless $d > 1$, in which case it converges. For (b) the easiest way to proceed is to change variables, with $u = 1 + e^x$ (it's more convenient to do this than $u = e^x$). ***For problems like this, look at the integrand as*** $x \to \pm\infty, 0$ ***and any other special points. If the integrand is not decaying at an appropriate rate....***

---

3. HW #3

Homework: Due Thursday October 1 (though you may place in my mailbox anytime up till 10am on Friday 10/2): Section 2.5: #2, #6. Section 2.7: #1, #4af, #7, #11, #18. Create two homework problems and TeX them up. They may be on anything related to probability; the first one you must be able to solve (and include the solution in your write-up); for the second, it's fine not to be able to do it (feel free to include a problem whose solution you'd like to know). I will share the problems and solutions with the class.

Section 2.5: Problem 2: For $\mathbb{P}(X = x, Y = y)$, as $Y = 1 - X$ this is impossible unless $(x, y) = (0, 1)$ or $(1, 0)$. The answer is that $(1, 0)$ happens with probability p, $(0, 1)$ with probability $1 - p$ and $(0, 0)$ and $(1, 1)$ have probability 0. The analysis for the second part is similar.

Section 2.5: Problem 6: Interestingly, this is *not* a cumulative distribution function (CDF)! While every square has four right angles and two pairs of parallel sides, it is not the case that every quadrilateral with four right angles and two pairs of parallel sides is a square (i.e., rectangles exist). Lemma 5 on page 39 lists properties a CDF should have; however, just because something has those properties does not make it a CDF. To be a CDF, it must assign non-negative probabilities to rectangles. We can evaluate the probability that $a \le X \le b$ and $c \le Y \le d$ in terms of F by using Problem 4 of this section. If we use that for this problem for the region $1 \le X \le 2$ and $1 \le Y \le 2$, we find that this square is assigned a negative probability; thus this F cannot be a CDF.

Remark 3.1. *This exercise is important as it illustrates a common theme: intuition in one-dimension frequently does not transfer to higher dimensions. On page 29 (Section 2.1) we learn that a similar lemma characterizes CDFs in one-variable – such a simple characterization does not hold in two dimensions. A very nice challenge problem is to see what conditions do uniquely characterize which functions are CDFs in two and higher dimensions.*

Section 2.7: Problem 1: As the probability that the first head occurs on toss n is $(1-p)^{n-1}p$, we have

$$\mathbb{P}(X > m) \ = \ \sum_{n=m+1}^{\infty} (1-p)^{n-1}p \ = \ (1-p)^m p \sum_{k=0}^{\infty} (1-p)^k \ = \ \frac{(1-p)^m p}{1-(1-p)} \ = \ (1-p)^m.$$

The distribution function is just 1 minus this (by the law of total probability), or $1 - (1-p)^m$. Note we could have also calculated $\mathbb{P}(X > m)$ by evaluating $1 - \mathbb{P}(X \le m)$ and using the *finite* geometric series formula.

Section 2.7: Problem 4af: Note F is continuous. For (a) we have

$$\mathbb{P}(1/2 \le X \le 3/2) \ = \ F(3/2) - F(1/2) \ = \ 3/4 - 1/4 \ = \ 1/2.$$

For (f):

$$\mathbb{P}(Z \le z) \ = \ \mathbb{P}(\sqrt{X} \le z) \ = \ \mathbb{P}(X \le z^2) \ = \ F(z^2)$$

as X is non-negative.

Section 2.7: Problem 7: The first airline is overbooked if all 10 seats are filled, which happens with probability $(9/10)^10 \approx 0.348678$. The second is overbooked if 20 or 19 people show up, which happens with probability $\binom{20}{20}(9/10)^{20}(1/10)^0 + \binom{20}{19}(9/10)^{19}(1/10)^1 \approx$ 0.391747. Thus there is a higher probability the second plane is overbooked. For problems like this, it is worthwhile trying to get a feel for the answer. Imagine we have a nine trillion seats and sell 10 trillion tickets. We expect nine trillion people to show up, and there should be approximately equal probability that more or less than nine trillion show up. Thus, in the limit as the size tends to infinity, there should be about a 50% chance the plane is overbooked, which is greater than the 34.87%.

Section 2.7: Problem 11: I'll write this up later.

Section 2.7: Problem 18: For both problems, there are $\binom{64}{8}$ ways to place the pawns on the board (where we do not care about the order in which the pawns are placed). (a) There are 18 ways to have 8 pawns in a line (8 horizontal lines, 8 vertical and two diagonal). Thus the probability that the 8 pawns are in a line is $8/\binom{64}{8}$. (b) There are $n!$ ways to place n pawns on an $n \times n$ board such that each row and each column has exactly one pawn. To see this, each pawn has coordinates (i, j) with $i, j \in \{1, \dots, n\}$. The solution is obtained by putting the pawns in order by their first coordinate; their second coordinates are just a permutation of $\{1, \dots, n\}$, and there are $n!$ such permutations. This problem is useful in abstract algebra. Instead of a chessboard, we consider an $n \times n$ matrix with a 1 if there is a pawn in the square, and a 0 otherwise. These matrices are called permutation matrices, and form a group under matrix multiplication. Cayley's theorem says any finite group is isomorphic to a subgroup of these matrices. (Note: for the problem asked in the book, we take $n = 8$ and find the probability that no two are in the same row or column is just $8!/\binom{64}{8}$.)

---

4. HW #4

Due Thursday October 8 (though you may place in my mailbox anytime up till 10am on Friday 10/9): Section 3.1: #1ac (hint: famous sum), #3 and Section 4.1: #1b. Section 3.2: #1, #4 and Section 4.2: #1 (also do when F is uniform on [0,1] and K = .9); obviously your solution will depend on the unknown distribution F. Section 3.3: #1, #2, #7 and Section 4.3: #1a, #2.

Section 3.1: Problem 1ac: For (a), the function is clearly non-negative. Its sum is

$$\sum_{n=1}^{\infty} \frac{C}{2^n} \;=\; C\sum_{n=1}^{\infty} \frac{1}{2^n} \;=\; C\frac{\frac{1}{2}}{1-\frac{1}{2}} \;=\; C;$$

thus to be a density we must take $C = 1$. Here we used the geometric series formula starting not at $n = 0$ but at $n = 1$. For (c), we again have a non-negative function whose sum now is

$$\sum_{n=1}^{\infty} \frac{C}{n^2} \;=\; C\sum_{n=1}^{\infty} \frac{1}{n^2} \;=\; C\frac{\pi^2}{6};$$

we have discussed this sum previously in class, and while it is not apparent why that sum is $\pi^2/6$, it should be clear that it is finite. The reason is this is a p-series from calculus: $\sum_{n=1}^{\infty} 1/n^p$ converges if $p > 1$ and diverges if $p \le 1$. Thus there is *some* choice of C so that the sum is 1. To see this, we use the Intermediate Value Theorem (IVT). Note the sum is clearly a continuous function of C. If $C = 0$ the sum is zero, while if $C = 2$ the sum is at least 2. Thus there is some choice of C so that the sum is 1 by the IVT. As $C\pi^2/6 = 1$, we see we must take $C = 6/\pi^2$.

The reason this problem is so important is that it is an example of the **Theory of Normalization Constants**. Namely, we frequently have a non-negative function that has a finite sum or integral, and thus there must be some way to rescale it so that it sums or integrates to 1; in other words, we can make it a probability density. This frequently arises in my work in Random Matrix Theory. I have horrible expressions involving $N(N+1)/2$ variables (with $N \to \infty$ and the normalization constant is given by a horrendous formula. Instead of working with that, I can just study the integrand and get it for free. One nice application of this is a proof of Wallis' formula for π; for details, see my paper in the Monthly:

```
http://www.williams.edu/go/math/sjmiller/public_html/math/
                         papers/StatProofWallis_Final.pdf
```

Section 3.1: Problem 3: *A fair coin is tossed n times. Every coin that lands on heads is tossed again. What is the probability mass function for the number of heads after the second toss?*

We solve this problem two ways. The first is the 'natural' approach. It has the advantage of being a reasonable method to try, but leads to a very messy formula.

Our first solution uses conditional probability. Let's say we want to compute all the ways of having m heads on the second toss, with clearly $0 \le m \le n$. We can express this probability as

$$\sum_{k=m}^{n} \mathbb{P}(m \text{ heads on second toss}|k \text{ heads on first}) \cdot \mathbb{P}(k \text{ heads on first toss}).$$

Why? We must have tossed some number of heads on the first toss, which we denote by k. Clearly $k \ge m$ as otherwise we can't have m heads on the second. The answer is thus

$$\sum_{k=m}^{n} \binom{k}{m} p^m (1-p)^{k-m} \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

It is worth asking what would happen if we forgot about the restriction that $m < n$; for example, what if $n = 4$ and $m = 6$? We would have the binomial coefficient $\binom{4}{6}$ – how is this defined? We might at first expect it to be $\frac{4!}{6!}(-2)!$; this works but you need to know that $(-2)!$ is defined to be infinity! We'll discuss this later when we talk about the Gamma function, which generalizes the factorial function. There is another way to 'see' what the definition should be. We expect the answer to be zero, as the combinatorial interpretation is: *how many ways are there to choose 6 objects from 4 when order doesn't matter?* Clearly there are *no* such ways, and thus the answer should be zero. Another way of defining $\binom{n}{k}$ is

$$\frac{n(n-1)\cdots(n-(k-1))}{k(k-1)\cdots 1}.$$

In our case, we would have

$$\binom{4}{6} \ = \ \frac{4\cdot 3\cdot 2\cdot 1\cdot 0\cdot(-1)}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} \ = \ 0$$

as we have a 0 in the numerator.

Remember, in mathematics we can make almost any definition we want – the question is when our definition is useful. The above is a great way to define the choose function when the bottom exceeds the top, and agrees with our combinatorial intuition.

We now give an alternate solution. A much better way to look at this problem is to think what must happen for a coin to end up heads after two tosses. The only way this can occur is if the first and second tosses are heads, which (since the coin lands on heads with probability p) happens with probability $p \cdot p = p^2$. Our situation turns out to be equivalent to the following: *Toss a biased coin (with probability p^2 of landing on heads) a total of n times; what is the probability mass function?* The answer is just

$$\mathbb{P}(m \text{ heads}) \ = \ \binom{n}{m}\left(p^2\right)^m\left(1-p^2\right)^{n-m} \ = \ \binom{n}{m}p^{2m}(1-p^2)^{n-m}.$$

The above analysis illustrates one of the most common ways to prove combinatorial identities. Namely, we calculate a given quantity two different ways. As both count the same object, they must be equal. Typically one is easily computed, and thus the other, harder combinatorial expression must equal the easier one. For example, in our case above the second approach was fairly easy to compute. If we take $p = 1/2$ and set the first and second solutions equal to each other, we find

$$\sum_{k=m}^{n}\binom{k}{m}\binom{n}{k}\left(\frac{1}{2}\right)^{n+k} \ = \ \binom{n}{m}\frac{3^{n-m}}{2^{2n}}.$$

We can verify this identity for any choices of $m \le n$; however, is there a way of proving this directly (and not relying on us being clever and noticing this counting problem was equivalent to another)?

Section 4.1: Problem 1b: Our proposed density is again non-negative, so the question is just whether or not it will integrate to 1 for some choice of C. We have

$$\int_{-\infty}^{\infty} C\exp(-x-\exp(-x))dx \ = \ C\int_{-\infty}^{\infty}\exp(-x)\exp(-\exp(-x))dx.$$

We do a u substitution. Let

$$u \ = \ \exp(-\exp(-x))$$

so

$$du \ = \ \exp(-x)\exp(-\exp(-x))dx,$$

and $x : -\infty \to \infty$ becomes $u : 0 \to 1$. Thus our integral is

$$C\int_0^1 du \ = \ 1.$$

There are other change of variables we could make, but this is the simplest. See the comments for Section 3.1, #1 for more on problems like this.

Section 3.2: Problem 1: Clearly all three are not independent, as if we know X and Y then we know Z. From construction, X and Y are independent, and by symmetry it suffices to show X and Z are independent (Y and Z are independent by a similar argument). To see that they are independent we must show

$$\mathbb{P}X = x, Z = z \ = \ \mathbb{P}X = x\mathbb{P}Z = z.$$

We have four possibilities: $x \in \{-1, 1\}$ and $z \in \{-1, 1\}$. A straightforward calculation shows each $\mathbb{P}X = x$ and $\mathbb{P}Z = z = 1/2$, while $\mathbb{P}X = x, Z = z = 1/4$.

Section 3.2: Problem 4: We use the idea from the basketball game in class, namely that this is a memoryless game. For the first problem, after A throws a 6 we do not care if she (A is obviously named Alice) throws another 6 before B (clearly Bob) or C (surely Charlie) does; all we care about is that B then throws a 6 before Charlie. Let x be the probability that A rolls the first 6. Then

$$x \ = \ \frac{1}{6} + \left(\frac{5}{6}\right)^3 x;$$

this is because she either rolls a 6 on her first try, or she and B and C all miss, and then it is as if we've started the game fresh. (Note how important the memoryless feature is in solving these problems!) We thus find $x = \frac{1}{6} + \frac{125}{216}x$, or after some algebra $x = \frac{36}{91}$. We now keep rolling, and we only care about the rolls of B and C. It suffices to determine the probability B gets the next 6, as clearly C will then be the last to roll (from a previous homework problem, related to Murphy's law, we do know eventually C will roll a 6). Let y be the probability B rolls a 6 before C, given that B rolls first. A similar analysis gives

$$y \ = \ \frac{1}{6} + \left(\frac{5}{6}\right)^2 y,$$

or $y = \frac{1}{6} + \frac{25}{36}y$, which gives $y = \frac{6}{11}$. Thus the probability that A is first, then B and then C is just

$$\frac{36}{91} \cdot \frac{6}{11} \cdot 1 \ = \ \frac{216}{1001}.$$

For the second part, we now want A to roll the first 6, and then the next 6 *must* be rolled by B, and then the next *must* be rolled by C; thus, we now care about A's subsequent rolls. Fortunately we've already solved this problem! In the analysis above, we may interpret $x = 36/91$ as the probability that the first 6 is rolled by the person currently rolling. Thus the answer here is just $x^3 = (36/91)^3$; the reason is that once A rolls a six, it is now B's turn to roll.

Section 4.2: Problem 1: This is another example of the geometric series / waiting for a success. The probability p that we have an acceptable offer is $1 - F(K)$, while the probability we have an offer that is too low is $1 - p = F(K)$. Thus the probability that the first acceptable offer is the n^{th} is just

$$(1-p)^{n-1}p \ = \ F(K)^{n-1}(1 - F(K)),$$

and hence the expected value is

$$\sum_{n=1}^{\infty} n(1-p)^{n-1}p \ = \ \sum_{n=1}^{\infty} F(K)^{n-1}(1-F(K).$$

From class we know that this sum is just $1/p$ (we proved this by differentiating the geometric series), and since $p = 1 - F(K)$ thus the expected value is

$$\frac{1}{1-F(K)} \ = \ \frac{1}{F(K)}.$$

If we take our density to be the uniform distribution on $[0,1]$ and $K = .9$, then $F(K) = .9$ and the answer is just 10. Note that we really don't need to know the density and K; all we need to know is the value of $F(K)$.

It is worth reflecting on whether or not an answer of $1/p$ is reasonable. If $p = 1$ then the expected value is 1 – this is eminently reasonable, as we clearly win on the first offer. If $p = 0$ then we never win, and this is seen by the expected number becoming infinite. It is always worth checking answers at extreme cases (or in limits as we approach extreme cases, such as $p \to 0$) to get some feel for what is going on.

Section 3.3: Problem 1: Usually $\mathbb{E}[1/X]$ is not $1/\mathbb{E}[X]$. Almost anything is a counter-example. A trivial one is to take $X = \pm 1$ with probability 1/2 for each. Another example is to take $X = 2$ or 4 with probability 1/2 for each, as

$$\mathbb{E}[1/X] \ = \ \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{4}\cdot\frac{1}{2} \ = \ \frac{3}{8},$$

while

$$\frac{1}{\mathbb{E}[X]} \ = \ \frac{1}{2\cdot\frac{1}{2} + 4\cdot\frac{1}{2}} \ = \ \frac{1}{3}.$$

It is possible for them to be equal – this is always the case if $X = x$ with probability 1 for some non-zero x. Assume we have $X = x_i$ with probability p_i for $i \in \{1,2\}$ and we want these two to be equal. As $p_2 = 1 - p_1$, letting $p = p_1$ that requires

$$\frac{p}{x_1} + \frac{1-p}{x_2} \ = \ \frac{1}{x_1 p + x_2(1-p)}$$

or

$$\frac{x_1(1-p) + p x_2}{x_1 x_2} \ = \ \frac{1}{x_1 p + x_2(1-p)},$$

which simplifies to

$$\left(x_1(1-p) + p x_2\right)\left(x_1 p + x_2(1-p)\right) - x_1 x_2 \ = \ 0.$$

Are there any non-trivial solutions to this? We have three unknowns and only one equation, so this should be solvable. Of course, we do have restrictions: $0 < p < 1$ and $x_1 \neq x_2$. (We take $p \neq 0, 1$ as otherwise this reduces to the trivial solution.)

Section 3.3: Problem 2: This is a beautiful problem illustrating the power of expectation. Not surprisingly, it starts off as another geometric series problem (i.e., waiting for the first success). Let Y_j be the random variable which denotes how much time we need to wait to get the next new coupon given that we have j distinct coupons (of the c coupons). For each pick, the probability we get one of the j coupons we already have is $\frac{j}{c}$, and thus the probability p we get a new coupon is $p = 1 - \frac{j}{c} = \frac{c-j}{c}$. Thus, letting $p = \frac{c-j}{c}$ we find the probability that we get the next new coupon on pick n is just $(1-p)^{n-1}p$, so the expected value is

$$\sum_{n=1}^{\infty} n \cdot (1-p)^{n-1}p \;=\; \sum_{n=1}^{\infty} \left(\frac{j}{c}\right)^{n-1} \frac{c-j}{c};$$

as $p = \frac{c-j}{c}$ and the expected value is $1/p$, we have $\mathbb{E}[Y_j] = \frac{c}{c-j}$. Note the answer is reasonable. When $j = 0$ the expected wait is just one pick (which makes sense, as we have no coupons so anything is new). When $j = c-1$ we are missing only one coupon, and the answer is an expected wait of c (also reasonable!).

For the second part, if Y is the random variable which denotes how long we must wait to get all the coupons, then $Y = Y_0 + \cdots + Y_{c-1}$. As expectation is linear,

$$\mathbb{E}[Y] \;=\; \mathbb{E}[Y_0] + \cdots + \mathbb{E}[Y_{c-1}] \;=\; \frac{c}{c-0} + \cdots + \frac{c}{c-(c-1)}.$$

If we read the sum in reverse order and factor out a c, we notice it is

$$\mathbb{E}[Y] \;=\; c\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{c}\right) \;\approx\; c\log c,$$

as the sum is the c^{th} harmonic number H_c, which is about $\log c$ (a better approximation is $\log c + \gamma$, where γ is the Euler-Mascheroni constant and is about .5772156649). See

`http://en.wikipedia.org/wiki/Harmonic_number`

for more information.

Section 3.3: Problem 7: First off, an A+ in the course to anyone who can find a real world example of this in time for us to make bets (note this may fail if there are administrative fees for placing bets). The problem means that if we place \$1 on horse i and that horse wins then we win $\pi(i)$ dollars and get to keep our initial wager; if that horse loses then we have lost our wager. Let us bet b_i dollars on horse i. Our total wager is $b_1 + \cdots + b_n$. If horse i wins then we win $\pi(i)b_i$ dollars and get to keep our wager of b_i; however, we have lost our wager everywhere else. The amount we've lost is clearly at most $b_1 + \cdots + b_n$ (we shouldn't include the b_i here, but it is easier to do so). Thus, as long as

$$(\pi(i)+1)b_i \;>\; \sum_{k=1}^{n} b_k$$

then we end with more money than we started, and thus we win. We are told that $\sum_{k=1}^{n} \frac{1}{\pi(i)+1} < 1$. If we let $b_i = \frac{1}{\pi(i)+1}$ then the sum of our bets is less than \$1, but if horse i wins we end with $(\pi(i)+1) \cdot \frac{1}{\pi(i)+1} = 1$. Thus, no matter which horse wins, we end up with more money than we started! What is truly amazing about this problem is that we *do not* need to know the probabilities of the horses winning! Another interesting point to note is that the amount we win is independent of which horse triumphs!

What if we knew that horse 3 had a 99.9% chance of winning, and everything else was as before. How should we place our bets? If we want to do math, exactly as before! The odds of a horse winning are immaterial for this analysis. If we knew that horse 3 would always win, yes, place all of our money on horse 3. If, however, we deviate from the derived betting distribution, we are no longer doing math but actually gambling.

A few years ago my brother and some of his colleagues had the following odds on the Patriots going undefeated in the regular season: Ed got 5 to 1, Bozo (aka, Jason) gave 8 to 1; Mike got 6 to 1, Bozo gave 8 to 1; Wilhelm got 5 to 1 and Bozo gave 7 to 1. Was my brother happy, and if so, why? How should he have bet?

Section 4.3: Problem 1a: We need

$$\int_0^\infty x^\alpha e^{-x} dx$$

to be finite. As the exponential function decays much faster than polynomials grow ($x^\alpha < e^{x/2}$ for x large), there is no trouble at infinity. We just need the integral to be well-defined near 0. Near 0, e^{-x} looks like 1, so we need x^α to be integrable near the origin. This forces $\alpha > -1$. To see this, note for $\alpha \neq -1$ we have

$$\lim_{\epsilon\to 0}\int_\epsilon^1 x^\alpha dx \ = \ \lim_{\epsilon\to 0} \frac{x^{1+\alpha}}{1+\alpha}\Big|_\epsilon^1 \ = \ \lim_{\epsilon\to 0}\frac{1-\epsilon^{1+\alpha}}{1+\alpha},$$

and this forces $\alpha > -1$. If $\alpha = -1$ then $\int dx/x$ is just $\ln x$, which blows up.

Section 4.3: Problem 2: This is one of my favorite problems. At first the answer seems too good to be true, as it is independent of the distribution of the X_i's! All that matters is that they are identically distributed and that the sum is non-zero (so the division makes sense). Let X have the same distribution as the X_i's. The key technique here is to multiply by 1. We start with

$$\mathbb{E}\left[\frac{1}{1}\right] \ = \ 1;$$

this trivial observation is the key to the proof. We now write 1/1 in a clever way, and use linearity of expectation:

$$\begin{aligned} 1 \ &= \ \mathbb{E}\left[\frac{X_1+\cdots+X_n}{X_1+\cdots+X_n}\right] \\ &= \ \sum_{k=1}^n \mathbb{E}\left[\frac{X_k}{X_1+\cdots+X_n}\right] \\ &= \ n\mathbb{E}\left[\frac{X}{X_1+\cdots+X_n}\right], \end{aligned}$$

and so

$$\mathbb{E}\left[\frac{X}{X_1+\cdots+X_n}\right] \ = \ \mathbb{E}\left[\frac{X_k}{X_1+\cdots+X_n}\right] \ = \ \frac{1}{n}.$$

The key step above is that as the X_k's are identically distributed, the expected value of any one of them over the sum is the same as that of any other over the sum. We now calculate the quantity of interest:

$$\mathbb{E}\left[\frac{X_1+\cdots+X_m}{X_1+\cdots+X_n}\right] \ = \ \sum_{k=1}^m \mathbb{E}\left[\frac{X_k}{X_1+\cdots+X_n}\right] \ = \ \frac{m}{n}.$$

---

5. HW #5

Due Thursday October 15 (though you may place in my mailbox anytime up till 10am on Friday 10/16): (1) Calculate the second and third centered moments of Binomial(n,p); (2) Calculate the kth centered moment of the standard normal. Section 3.4: #1. Section 3.11: #9, #13 (can do after Tuesday's lecture). Section 4.14: #12.

First Problem: *Calculate the second and the third moments of X when $X \sim \mathrm{Bin}(n, p)$ (this means X is a random variable with the binomial distribution with parameters n and p.*

One natural way to compute this is from the definition. To evaluate the second moment, we either need to compute $\mathbb{E}[(X-\mu)^2]$ or $\mathbb{E}[X^2]-\mathbb{E}[X]^2$. In the latter, this leads us to finding

$$\sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

While we can do this through differentiating identities, it is faster to use linearity of expectation. Let $X_1, \dots, X_n$ be i.i.d.r.v. (independent identically distributed random variables) with the Bernoulli distribution with parameter p. Note these are independent, and we have the probability X_i is 1 is p and the probability X_i is 0 is $1-p$. Let $X = X_1 + \cdots + X_n$. As they are independent, the variance of the sum is the sum of the variances:

$$\mathrm{Var}(X_1 + \cdots + X_n) \ = \ \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \ = \ np(1-p),$$

as the variance of each X_i is just $p(1-p)$. To see this, note

$$\mathbb{E}[(X_i - \mu_i)^2] \ = \ \mathbb{E}[(X_i - p)^2] \ = \ (1-p)^2 p + (0-p)^2 p \ = \ p(1-p).$$

We redo the calculations in a way that will help with the analysis of the third moment. We have

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[(X_1 + \cdots + X_n)^2] \\
&= \mathbb{E}[X_1^2 + \cdots + X_n^2 + 2X_1X_2 + 2X_2X_3 + \cdots + 2X_{n-1}X_n] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[X_iX_j].
\end{aligned}$$

As the X's are independent, $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = p^2$ (so long as $i \neq j$); note there are $\binom{n}{2}$ pairs (i,j) with $1 \le i < j \le n$. What about $\mathbb{E}[X_i^2]$? That is readily seen to be just $1^2 \cdot p + 0^2 \cdot (1-p) = p$. Substituting gives

$$\mathbb{E}[X^2] \ = \ \sum_{i=1}^{n} p + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p^2 \ = \ np + \binom{n}{2} p^2.$$

Thus the variance is

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 \quad = \quad np + 2\frac{n(n-1)p^2}{2} - (np)^2 \; = \; np - np^2 \; = \; np(1-p).$$

We thus recover our result from above.

How should we handle the third moment? As $\mathbb{E}[X] = np$ and $\mathbb{E}[X^2] = p$, we have

$$\begin{aligned}
\mathbb{E}[(X-\mu)^3] &= \mathbb{E}[X^3 - 3X^2\mu + 3X\mu^2 - \mu^3] \\
&= \mathbb{E}[X^3] - 3np\mathbb{E}[X^2] + 3(np)^2\mathbb{E}[X] - (np)^3 \\
&= \mathbb{E}[X^3] - 3n^2p^2(1-p) + 3n^3p^3 - n^3p^3.
\end{aligned}$$

We can complete the analysis in a similar manner as above, namely expanding out

$$X^3 \; = \; (X_1 + \cdots + X_n)^3$$

and then using linearity of expectation. At this point, differentiating identities isn't looking so bad!

To solve this with differentiating identities, we must evaluate a sum such as

$$\sum_{k=0}^{n} k^3 \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

We start with the identity

$$(x+y)^n \; = \; \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

We apply the operator $x\frac{d}{dx}$ three times to each side, and find (after some tedious but straightforward algebra and calculus) that the left hand side equals

$$nx(x+y)^{n-3}\left(n^2x^2 + 3nxy - y(x-y)\right).$$

Setting $y = 1 - x$ and $x = p$ yields

$$np\left(1 + 3(n-1)p + (n^2 - 3n + 2)p^2\right) \; = \; \sum_{k=0}^{n} k^3 \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

The above is quite messy, and there is a very good chance we have made an algebra mistake. Thus, let's see if we can find another approach which will lead to cleaner algebra. Instead of applying $x\frac{d}{dx}$ three times, let's apply $x^3\frac{d}{dx}$. Applying this to $(x+y)^n$ is very easy, giving $x^3 \cdot n(n-1)(n-2)(x+y)^{n-3}$; applying it to the combinatorial expansion gives not k^3 and $k(k-1)(k-2)$. Collecting, we find

$$\begin{aligned} n(n-1)(n-2)x^3(x+y)^{n-3} &= x^3 \sum_{k=0}^{n} k(k-1)(k-2)\binom{n}{k} x^{k-3} y^{n-k} \\ &= \sum_{k=0}^{n} \left(k^3 - 3k^2 + 2k\right) \binom{n}{k} x^k y^{n-k} \\ &= \sum_{k=0}^{n} k^3 \binom{n}{k} x^k y^{n-k} - 3\sum_{k=0}^{n} k^2 \binom{n}{k} x^k y^{n-k} \\ &\qquad + 2\sum_{k=0}^{n} k \binom{n}{k} x^k y^{n-k}. \end{aligned}$$

Setting $x = p$ and$y = 1 - p$ yields

$$n(n-1)(n-2)p^3 \;=\; \mathbb{E}[X^3] - 3\mathbb{E}[X^2] + 2\mathbb{E}[X].$$

We have made a lot of progress, as we already know $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ and can thus solve for $\mathbb{E}[X^3]$. The point is that it is easier *not* to try and find $\mathbb{E}[X]$ directly, but rather to find a related quantity. Note, of course, that this method requires us to know $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ before we can deduce the value of $\mathbb{E}[X^3]$; this is not an unreasonable request, as typically we want to know all the moments up to a certain point.

The general principle here is that algebra can be hard, painful and tedious, but if you look at a problem the right way, you can minimize how much algebra you need to do. It's worthwhile to spend a few minutes thinking about how we can try and approach a problem, as often this leads to a way with significantly less messy computations.

Second Problem: *Calculate the k^{th} moment of the standard normal.* The density function of the standard normal is $(2\pi)^{-1/2}\exp(-x^2/2)$. We are thus reduced to calculating

$$M(k) \;=\; \int_{-\infty}^{\infty} x^k \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

The integral is clearly zero for k odd, as we are integrating an odd function over a symmetric region. (Note the normal decays so rapidly that all the integrals exist). There are at least two natural ways to handle even k.

The standard approach is through induction and integration by parts. Consider

$$\int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

To integrate by parts, we need to choose values for u and dv. While at first we might think the natural choices are either $u = x^2$ or $dv = x^2$, if we try either we run into problems. The reason is that there is no nice anti-derivative for $e^{-x^2/2}$. Fortunately, all is not lost. The function $e^{-x^2/2}$ is *screaming* to us that it wants to be considered with a factor of x, as then it *will* have a nice anti-derivative. Thus we try

$$u \;=\; x, \quad dv \;=\; \frac{1}{\sqrt{2\pi}} e^{-x^2/2} x dx.$$

This leads to $du = dx$ and $v = -(2\pi)^{-1/2}e^{-x^2/2}$. Thus we find

$$M(2) \ = \ uv\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \ = \ I(0) \ = \ 1.$$

We have thus shown that the second moment is 1!

More generally, assume we know $M(2k) = (2k-1)!!$. Then we proceed as above, and to compute $M(2k+2)$ when we integrate by parts we set $u = x^{2k+1}$, so $du = (2k+1)x^{2k}dx$. The boundary term vanishes when evaluated at $\pm\infty$, and we find

$$\begin{aligned} M(2k+2) \ &= \ (2k+1)\int_{-\infty}^{\infty} x^{2k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \ (2k+1)M(2k) \ = \ (2k+1)(2k-1)!! \ = \ (2k+1)!!. \end{aligned}$$

Similar to the previous problem, we show how it may also be done through differentiating identities. It seems strange to talk about differentiating identities here, as

$$I(2k) \ = \ \int_{-\infty}^{\infty} x^{2k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

has no free parameter! We begin with the fact that

$$1 \ = \ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx;$$

this is just the statement that the above is the probability density for a normal distribution with mean 0 and variance σ^2. Moving σ to the other side gives

$$\sigma \ = \ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2} dx.$$

We keep applying $\sigma^3 \frac{d}{d\sigma}$ to both sides/ Why do we multiply by σ^3? The reason is that the differentiation hits $-x^2\sigma^{-2}/2$, and thus brings down a factor of $x^2\sigma^{-3}$. Hence if we multiply by σ^3, we keep everything nice. Differentiating once gives

$$\sigma^3 \cdot 1 \ = \ I(2).$$

Applying $\sigma^3 \frac{d}{d\sigma}$ again gives

$$\sigma^3 \cdot (3 \cdot 1\sigma^2) \ = \ I(4),$$

or $3!!\sigma^5 = I(4)$. Differentiating again gives $5!!\sigma^7 = I(6)$, and by induction we can show $(2k-1)!!\sigma^{2k+1} = I(2k)$. Setting $\sigma = 1$ completes the analysis.

Note for this problem that while differentiating identities is quite useful, it was not immediately apparent what identity we needed to use!

Section 3.4: Problem 1: We toss a coin that lands heads p percent of the time a total of n times, and want to know the expected number and variance of the number of runs. Remember a run is a set of consecutive heads or tails, and when we get a coin of the opposite value then we start a new run. For example, in $HHTTTTHTHT$ we have 6 runs; we start with a run of two heads, then have a run of four tails, then a run of one head followed by a run of one tail followed by a run of one head followed by a run of one tail.

We solve the problem using binary indicator random variables and expectation. For $i \in \{2, \dots, n\}$, let $X_i = 1$ if toss i is different than toss $i-1$, and 0 if the two tosses are the same. Note the X_i's are all identically distributed Bernoulli random variables with probability $2p(1-p)$. (To see this is the probability, note we just have to figure out how likely it is to get HT or TH.) We let

$$X \ = \ 1 + X_2 + \cdots + X_n;$$

we start with a 1 as the first toss always starts a run. Using linearity of expectation,

$$\mathbb{E}[X] \ = \ 1 + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] \ = \ 1 + (n-1)2p(1-p).$$

Whenever we prove a formula, it is always worthwhile to see if it is reasonable. If $p = 0$ or $p = 1$ then there is only one run, as expected. Further, note the expected number of runs is largest when $p = 1/2$ (this is a nice calculus problem, namely showing that the maximum value of $p(1-p)$ happens when $p = 1/2$). In this case we get $\frac{n+1}{2}$ runs, which is the average of n (the most runs we could have) and 1 (the fewest number of runs we could have).

How do we compute the variance? Note that we may ignore the $+1$ term and just study $X_2 + \cdots + X_n$. We have the formula

$$\text{Var}(X_2 + \cdots + X_n) \ = \ \sum_{i=2}^{n} \text{Var}(X_i) + 2\sum_{i=2}^{n-1}\sum_{j=i+1}^{n} \text{CoVar}(X_i, X_j).$$

For notational convenience, set $q = 2p(1-p)$, and note that each X_i is a Bernoulli random variable with parameter q. The variance of each X_i is therefore just $q(1-q)$. What about the covariance terms? We have

$$\text{CoVar}(X_i, X_j) \ = \ \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

If $j \ge i+2$ then X_j and X_i are independent (and thus there covariance is zero), while if $j = i+1$ they are dependent. In this latter case, we have $\mathbb{E}[X_i]\mathbb{E}[X_j] = q^2$, while

$$\mathbb{E}[X_iX_j] \ = \ 1\cdot\left(p^2(1-p) + p(1-p)^2\right) + 0\cdot\left(1 - \left(p^2(1-p) + p(1-p)^2\right)\right) \ = \ p(1-p);$$

the reason this is the answer is that the only way for $X_iX_j = 1$ when $j = i+1$ is for us to have HTH or THT. Therefore

$$\text{Var}(X_2 + \cdots + X_n) \ = \ (n-1)q(1-q) + (n-1)\left(p(1-p) - q^2\right),$$

with $q = 2p(1-p)$. Simple algebra shows $p(1-p) = \frac{q}{2}$; as $q \le 1/2$, $p(1-p) \ge q^2$. Thus our variance is non-negative, and is a constant times n, which implies our answer is 'reasonable'.

Section 3.11: Problem 9: We let X represent the number of heads in n tosses of a biased coin that is heads with probability p (note we are changing notation slightly from the book). Thus

$$\mathbb{P}(X = k) \ = \ \binom{n}{k} p^k (1-p)^{n-k}.$$

We wish to compute the probability that X is even; thus we need to evaluate

$$\sum_{i=0}^{n/2} \binom{n}{2i} p^{2i}(1-p)^{n-2i}.$$

An elegant way to solve this is to consider

$$\frac{1}{2}(x+y)^n + \frac{1}{2}(y-x)^n.$$

When we expand this out, only the terms involving x to an even power survive. Setting $x = p$ and $y = 1-p$ yields

$$\frac{1}{2}1^n + \frac{1}{2}(1-2p)^n \ = \ \sum_{i=0}^{n/2} \binom{n}{2i} p^{2i}(1-p)^{n-2i},$$

and thus the probability that X is even is $\frac{1}{2} + \frac{(1-2p)^n}{2}$.

As always, our first thought should be: is our answer reasonable? As $-1 \le 1-2p \le 1$, we see our probability is always between 0 and 1. For our next test, it is good to consider extreme cases. What happens if $p = 0$? Then there are never any heads and, as zero is an even number, we should (and do!) have an even number of heads with probability 1. If instead $p = 1$ then there are an even number of heads if n is even else there is an odd number of heads; both of these observations are satisfied by our answer. Finally, if $p = 1/2$ then there is precisely a 50% chance of having an even number of heads. Is this reasonable? *YES!* To see why this is reasonable, note that it doesn't matter what the first $n-1$ tosses are; given any outcomes there, we have a 50% chance that we have an even number of heads after the last toss (if there is already an even number of heads we need a tail, while if there is an odd number of heads then we need a head, with each of these events happening with probability one-half).

While this is an interesting problem, to me the really important aspect is seeing whether or not our answer at the end of the day is reasonable. Learning how to do these quick tests / checks is a very important skill.

Section 3.11: Problem 13: I have discussed this with a few of you. In the interest of time, I'm hoping to modify someone's TeX code here. The key observation is that we need to use a generalization of the cookie problem.

Section 4.14: Problem 12: A random variable has a chi-square distribution with d degrees of freedom if it has density

$$f_d(x) \ = \ \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x/2} & \text{if } x \ge 0 \\ 0 & \text{otherwise,} \end{cases}$$

where Γ is the Gamma function (the generalization of the factorial function), which is given by

$$\Gamma(s) \ = \ \int_0^\infty x^{s-1}e^{-x}dx.$$

We first show that if $X_1 \sim N(0,1)$ (which means X_1 is normally distributed with mean 0 and variance 1) then X_1^2 is a chi-square distribution with 1 degree of freedom.

Let $Y = X_1^2$. Then

$$\begin{aligned} \mathbb{P}(Y \le y) &= \mathbb{P}(X^2 \le y) \\ &= \mathbb{P}\left(-\sqrt{y} \le X \le \sqrt{y}\right) \\ &= F(\sqrt{y}) - F(-\sqrt{y}), \end{aligned}$$

where F is the anti-derivative of the standard normal's density $f(x) = (2\pi)^{-1/2}e^{-x^2/2}$. At first it looks like we have made no progress, as there is no nice, closed form expression for the anti-derivative of the standard normal. All is not lost, however. The reasons this is progress is that the derivative of the cumulative distribution function is the density. Thus, the density of Y, which we denote by $h(y)$, is given by the derivative of $\mathbb{P}(Y \le y)$ with respect to y. Using the chain rule, we find

$$\begin{aligned} h(y) &= \frac{d}{dy}\left[F(\sqrt{y}) - F(-\sqrt{y})\right] \\ &= F'(\sqrt{y})\frac{1}{2\sqrt{y}} - F'(-\sqrt{y})\frac{-1}{2\sqrt{y}} \\ &= f(\sqrt{y})y^{-1/2} \ = \ \frac{1}{\sqrt{2\pi}}e^{-y/2}y^{\frac{1}{2}-1}, \end{aligned}$$

which is the density of a chi-square distribution with 1 degree of freedom (we need the fact that $\Gamma(1/2) = \sqrt{\pi}$).

What about the sum of the squares of two independent standard normal distributions? We again calculate the cumulative distribution function and then differentiate. We find

$$\begin{aligned} H(y) &= \mathbb{P}(X_1^2 + X_2^2 \le y) \\ &= \int\int_{x_1^2+x_2^2 \le y} \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}dx_1dx_2 \\ &= \int\int_{x_1^2+x_2^2 \le y} \frac{1}{2\pi}e^{-(x_1^2+x_2^2)/2}dx_1dx_2. \end{aligned}$$

We now switch to polar coordinates, setting $x_1 = r\cos\theta_1$ and $x_2 = r\sin\theta_2$. The change of variables formula gives $dx_1dx_2 = rdrd\theta$, and we obtain

$$\begin{aligned} H(y) &= \int_{\theta=0}^{2\pi}\int_{r=0}^{\sqrt{y}} \frac{1}{2\pi}e^{-r^2/2}rdrd\theta \\ &= \int_{r=0}^{\sqrt{y}} e^{-r^2/2}rdr \\ &= 1 - e^{-y/2} \end{aligned}$$

(the integration is up to $\sqrt{y}$ and not y as the radius-squared is y). Now that we know the cumulative distribution function $H(y)$, the density is simply the derivative. Thus we finally obtain

$$h(y) \ = \ \frac{1}{2}e^{-y/2},$$

which by inspection is the density of the chi-square distribution with two degrees of freedom.

Given the amount of work it took to evaluate the sum of the squares of two standard normal distributions, we are justified in being a little afraid of the calculation for the sum of n squares. It seems like we will need to know the change of variable formula for n-dimensional cartesian coordinates to n-dimensional spherical coordinates! Amazingly, though there are nice formulas for this, we do not need to know them because we will exploit a method known as the **Theory of Normalization Constants**. We know we can represent $x_1, \dots, x_k$ through the radius r and $k-1$ angles $\theta_1, \dots, \theta_k$. We have relations of the form

$$\begin{aligned} x_1 &= rg_1(\theta_1, \dots, \theta_{k-1}) \\ &\vdots \\ x_k &= rg_k(\theta_1, \dots, \theta_{k-1}). \end{aligned}$$

We state the Change of Variables Theorem:

Theorem 5.1 (Change of Variables)**.** *Let V and W be bounded open sets in $\mathbb{R}^k$. Let $h : V \to W$ be a 1-1 and onto map, given by*

$$h(u_1, \dots, u_k) \ = \ (h_1(u_1, \dots, u_k), \dots, h_k(u_1, \dots, u_k)). \tag{5.1}$$

Let $f : W \to \mathbb{R}$ be a continuous, bounded function. Then

$$\begin{aligned} &\int \cdots \int_W f(x_1, \dots, x_k) dx_1 \cdots dx_k \\ &\quad = \int \cdots \int_V f\left(h(u_1, \dots, u_k)\right) J(u_1, \dots, u_v) du_1 \cdots du_k, \end{aligned} \tag{5.2}$$

*where J is the **Jacobian***

$$J \ = \ \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \cdots & \frac{\partial h_1}{\partial u_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_k}{\partial u_1} & \cdots & \frac{\partial h_k}{\partial u_k} \end{vmatrix}. \tag{5.3}$$

If we are to use this theorem, we would need to compute the Jacobian, which would require us to know the change of variable functions f_i. Here is how we get around it. We need to figure out how the volume element $dx_1 \cdots dx_k$ changes; we clearly have

$$dx_1 \cdots dx_k \ = \ \mathcal{G}(r, \theta_1, \dots, \theta_{k-1}) dr d\theta_1 \cdots d\theta_{k-1}.$$

We must have

$$\mathcal{G}(r, \theta_1, \dots, \theta_{k-1}) \ = \ r^{k-1} \mathcal{C}(r, \theta_1, \dots, \theta_{k-1}).$$

Why? This follows from unit analysis. In two dimensions we have $dx_1 dx_2 \mapsto r dr d\theta$ and in three dimensions it is $dx_1 dx_2 dx_3 \mapsto r^2 \sin\theta_1 dr d\theta_1 d\theta_2$. Note that we have the radius to a power one less than the number of variables. This is because the angular variables are unitless, and thus the units of $dr d\theta_1 \cdots d\theta_{k-1}$ are meters (say), while $dx_1 \cdots dx_k$ has units of meters^k. Thus we need the factor r^{k-1}. We have therefore shown that there is some complicated function $\mathcal{C}$ such that

$$dx_1 \cdots dx_k \ = \ r^{k-1} \mathcal{C}(\theta_1, \dots, \theta_{k-1}) dr d\theta_1 \cdots d\theta_{k-1}.$$

We now return to our problem. Let $Y = X_1^2 + \cdots + X_k^2$. We again use the theory of cumulative distribution functions and find

$$\begin{aligned} H(Y) &= \mathbb{P}(X_1^2 + \cdots + X_k^2 \le y) \\ &= \int \cdots \int_{x_1^2+\cdots+x_k^2 \le y} \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} dx_1 \cdots dx_k \\ &= \int \cdots \int_{x_1^2+\cdots+x_k^2 \le y} \frac{1}{(2\pi)^{k/2}} e^{-(x_1^2+\cdots+x_k^2)/2} dx_1 \cdots dx_k. \end{aligned}$$

We now change variables. We don't care what the angular integrations are over, so we just denote those by ℓ_i to u_i (for lower and upper bound):

$$H(y) = \int_{r=0}^{\sqrt{y}} \int_{\theta_1=\ell_1}^{u_1} \cdots \int_{\theta_{k-1}=\ell_{k-1}}^{u_{k-1}} \frac{1}{(2\pi)^{k/2}} e^{-r^2/2} r^{k-1} \mathcal{C}(\theta_1, \dots, \theta_{k-1}) dr d\theta_1 \cdots d\theta_{k-1}.$$

We integrate over the $k-1$ angles; the answer is independent of r and y, and we denote it by C_k (it does depend on the number of angular variables). Hence

$$H(y) \ = \ C_k \int_{r=0}^{\sqrt{y}} e^{-r^2/2} r^{k-1} dr.$$

Let $f(r) = C_k e^{-r^2/2} r^{k-1}$ and $F(r)$ be its anti-derivative. Then

$$H(y) \ = \ F(\sqrt{y}) - F(0).$$

We take the derivative and finally (almost) obtain the density:

$$h(y) \ = \ F'(\sqrt{y}) \frac{1}{2\sqrt{y}} \ = \ \frac{C_k}{2} e^{-y/2} y^{\frac{k}{2}-1}.$$

Why do we say 'almost' above? The problem is we still have the constant C_k, which we should have determined by doing the angular integrations but did not. Thus we do not have the final answer; fortunately, it is trivial to compute C_k now. This seems absurd – how can we compute C_k now? Shouldn't we have computed it earlier? And, if we are going to compute it, shouldn't we figure out what the change of variable formulas are for going from Cartesian to spherical?

The reason we can evaluate it so easily is that $Y = X_1^2 + \cdots + X_k^2$ is a random variable; **therefore its density must integrate to 1!** We know from above the formula for the density of a chi-square random variable with k degrees of freedom; using y for the dummy variable it is just (for $y \ge 0$)

$$\frac{1}{2^{k/2}\Gamma(k/2)} y^{\frac{k}{2}-1} e^{-y/2}.$$

Note this has exactly the same y-dependence as our part, and thus the normalization constants must match up!

This is a very important problem, without a doubt the most important on this homework assignment. While there are other ways to compute this answer by doing more direct computations, I prefer this approach as it illustrates the power of the Theory of Normalization Constants. It's incredible how it allows us to bypass certain painful computations. This arises all the time in random matrix theory, one of my main research *interests.*

Remark 5.2. *If we hadn't been given the probability density function for a chi-square with k degrees of freedom, we could still have found the value of C_k by noting that $h(y)$ integrates to 1. We need to use the Gamma function, which is defined by*

$$\Gamma(s) \ = \ \int_0^\infty e^{-x} x^{s-1} dx.$$

Though we don't need it, it is worth noting for future problems that the Gamma function is a generalization of the factorial function. It's a nice exercise to prove $\Gamma(n+1) = n!$ for n a positive integer – the proof is by integrating by parts.

Returning to our problem, we have

$$\begin{aligned} 1 \ &= \ \int_0^\infty h(y) dy \\ &= \ \frac{C_k}{2} \int_0^\infty e^{-y/2} y^{\frac{k}{2}-1} dy. \end{aligned}$$

We change variables, letting $x = y/2$ so $dy = 2dx$ and find

$$1 \ = \ \frac{C_k}{2} \int_0^\infty e^{-x} 2^{\frac{k}{2}-1} x^{\frac{k}{2}-1} 2dx \ = \ \frac{C_k}{2} \cdot 2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right),$$

which implies

$$\frac{C_k}{2} \ = \ \frac{1}{2^{k/2}\Gamma(k/2)}.$$

Remark 5.3. *Whenever we see a new method, it's worth exploring how far we can push it. What else can we glean from the above analysis? Implicit in our computation is the 'surface area' of the n-dimensional sphere! Remember our volume element became*

$$r^{k-1} \mathcal{C}(\theta_1, \dots, \theta_{k-1}) dr d\theta_1 \cdots d\theta_{k-1},$$

and we showed

$$\int_{\theta_1=\ell_1}^{u_1} \cdots \int_{\theta_{k-1}=\ell_{k-1}}^{u_{k-1}} \frac{1}{(2\pi)^{k/2}} \mathcal{C}(\theta_1, \dots, \theta_{k-1}) d\theta_1 \cdots d\theta_{k-1} \ = \ C_k.$$

Using our value for C_k above, we find

$$\int_{\theta_1=\ell_1}^{u_1} \cdots \int_{\theta_{k-1}=\ell_{k-1}}^{u_{k-1}} \mathcal{C}(\theta_1, \dots, \theta_{k-1}) d\theta_1 \cdots d\theta_{k-1} \ = \ \frac{2(2\pi)^{k/2}}{2^{k/2}\Gamma(k/2)} \ = \ \frac{2 \cdot \pi^{k/2}}{\Gamma(k/2)}.$$

We claim that this is the surface area of the n-dimensional sphere. Why? We were integrating a function that depended only on the radius; thus we may consider our change of variables as partitioning the n-dimensional sphere of radius $\sqrt{y}$ into a collection of shells of radii ranging from 0 to $\sqrt{y}$. What does this formula give for specific n? We find

$$\begin{aligned} n = 2 \ &: \ 2\pi \\ n = 3 \ &: \ 4\pi \\ n = 4 \ &: \ 2\pi^2; \end{aligned}$$

except for the last, the previous two are well-known as the perimeter of the unit circle and the surface area of the unit sphere.

---

6. HW #6

Due Thursday October 22 (though you may place in my mailbox anytime up till 10am on Friday 10/16): Section 3.5: #2. Section 4.4: #5. Section 3.6: #2, #7. Also TeX up two problems; you must include an answer for the first.

Section 3.5: Problem 2: We toss N coins (each of which is heads with probability p), where $N \sim \text{Poisson}(\lambda)$, and let X denote the number of heads. What is the probability mass function of X? We compute it by calculating the probability of getting m heads when we toss n coins, and weight that by the probability of having n coins to toss. Thus the answer is

$$\begin{aligned}
\text{Prob}(X = m) &= \sum_{n=m}^{\infty} \text{Prob}(X = m|N = n) \cdot \text{Prob}(N = n) \\
&= \sum_{n=m}^{\infty} \binom{n}{m} p^m (1-p)^{n-m} \cdot \frac{\lambda^n e^{-\lambda}}{n!} \\
&= p^m e^{-\lambda} \sum_{n=m}^{\infty} \frac{n!}{m!(n-m)!} (1-p)^{n-m} \frac{\lambda^n}{n!} \\
&= \frac{p^m e^{-\lambda}}{m!} \sum_{n=m}^{\infty} \frac{(1-p)^{n-m} \lambda^n}{(n-m)!}.
\end{aligned}$$

We need to be ‘clever’ here to simplify the algebra and get a nice, clean expression, but note the very large hints. First off, we have a factor of $p^m e^{-\lambda}/m!$ outside. This looks a bit like the mass function of a Poisson, but not quite. Second, the sum above has two pieces that depend on $n - m$ and one piece that depends on n. This suggests we should add zero, and write

$$\lambda^n = \lambda^{n-m+m} = \lambda^{n-m} \cdot \lambda^m.$$

We can then pull the λ^m outside of the sum and we find

$$\text{Prob}(X = m) = \frac{p^m \lambda^m e^{-\lambda}}{m!} \sum_{n=m}^{\infty} \frac{(1-p)^{n-m} \lambda^{n-m}}{(n-m)!}.$$

We now let $k = n - m$ so the sum runs from 0 to ∞. We also combine the factors, and obtain

$$\begin{aligned}
\text{Prob}(X = m) &= \frac{(p\lambda)^m e^{-\lambda}}{m!} \sum_{k=0}^{\infty} \frac{((1-p)\lambda)^k}{k!} \\
&= \frac{(p\lambda)^m e^{-\lambda}}{m!} e^{(1-p)\lambda}
\end{aligned}$$

from the definition of e^x as

$$e^x \ = \ \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Simplifying the above expression, we finally obtain

$$\text{Prob}(X = m) \ = \ \frac{(p\lambda)^m e^{-p\lambda}}{m!},$$

which is the probability mass function for a Poisson random variable with parameter $p\lambda$.

It takes awhile to become proficient and fluent with such algebraic manipulations. A good guiding principle is that we want to manipulate the expressions towards some known end, which guides us in how to multiply by 1 or add 0. Here the key step was writing λ^n and $\lambda^{n-m}\lambda^m$.

As another example, let's compute the average value of a random variable Y with the Poisson distribution with parameter λ. We have

$$\begin{aligned}\mathbb{E}[Y] \ &= \ \sum_{n=0}^{\infty} n \cdot \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \ \sum_{n=1}^{\infty} n \cdot \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \ e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!}.\end{aligned}$$

To finish the evaluation, it is natural to write λ^n and $\lambda^{n-1}\lambda$. The reason for this is that we have a sum where the denominator involves $n-1$, and thus it is helpful to make the numerator depend on $n-1$ as well. If we let $k = n-1$, then as n runs from 1 to ∞ we have k runs from 0 to ∞, and we find

$$\mathbb{E}[Y] \ = \ e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \cdot \lambda}{k!} \ = \ \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \ = \ \lambda e^{-\lambda} e^{\lambda} \ = \ \lambda,$$

where again we made use of the series expansion of e^x.

Using this fact, we can find the expected number of heads in the assigned problem *without* actually proving that X is given by the Poisson distribution with parameter λp. To see this, we claim that if

$$\text{Prob}(X = m) \ = \ \sum_{n=m}^{\infty} \text{Prob}(X = m | N = n) \cdot \text{Prob}(N = n),$$

then

$$\mathbb{E}[X] \ = \ \sum_{n=0}^{\infty} \mathbb{E}[X | N = n] \cdot \text{Prob}(N = n),$$

which leads to

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=0}^{\infty} np \cdot \frac{\lambda^n e^{-\lambda}}{n!} \\ &= p\sum_{n=0}^{\infty} n \cdot \frac{\lambda^n e^{-\lambda}}{n!};\end{aligned}$$

the last sum is just the expected value of the Poisson distribution with parameter λ, which we know is λ. Thus $\mathbb{E}[X] = p\lambda$.

Section 4.4: Problem 5: We want to compute the density of $Y = e^X$, where $X \sim N(0,1)$. The latter means that X has the standard normal distribution, namely that the density function of X, f_X, satisfies

$$f_X(x) \ = \ \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}.$$

One very easy way to compute the answer to problems like this is by using cumulative distribution functions, and noting the probability density is the derivative. Let F_X and F_Y represent the cumulative distribution functions of X and Y, and let f_X and f_Y denote their densities. We have

$$\begin{aligned}F_Y(y) &= \text{Prob}(Y \le y) \\ &= \text{Prob}(e^X \le y) \\ &= \text{Prob}(X \le \log y) \\ &= F_X(\log y).\end{aligned}$$

We now differentiate, using the chain rule.

$$f_Y(y) \quad = \quad F_X'(\log y)\cdot(\log y)' \ = \ f_X(\log y)\cdot\frac{1}{y}.$$

Substituting for f_X, we obtain

$$f_Y(y) \ = \ \frac{1}{\sqrt{2\pi}}\,\frac{1}{y}\,e^{-\frac{\log^2(y)}{2}}.$$

Section 3.6: Problem 2: We are asked to find the marginal densities for a multinomial distribution with parameters n and $p_1, \dots, p_t$. Without loss of generality we may find the marginal for the last variable, as the other cases are handled analogously. Note that for a multinomial, we have $p_1 + \cdots + p_t = 1$, and $n = n_1 + \cdots + n_t$. Let X_t be the random variable for the last variable. If we want to calculate the probability that X_t equals m say, we must sum over all the remaining variables (I prefer to use a different letter for the variable of interest to emphasize that we do not wish to sum over it). As the sum of all t variables is n and the last variable is m, we are simply summing over $n_1 + \cdots + n_{t-1} = n - m$. We thus have

$$\text{Prob}(X_t = m) \quad = \quad \left(\sum_{n_1+\cdots+n_{t-1}=n-m} \frac{n!}{n_1!\cdots n_{t-1}!}p_1^{n_1}\cdots p_{t-1}^{n_{t-1}}\right)\frac{1}{m!}p_t^m.$$

The quantity in parentheses looks a lot like a multinomial sum with $n-1$ probabilities; it is not quite that, as the numerator is supposed to equal the sum of the numbers being factorialed in the denominator. This is readily fixed. We need the numerator to be $(n-m)!$ instead of $n!$, so we multiply by 1, replacing $n!$ with $(n-m)! \cdot n!/(n-m)!$. This leads us to

$$\begin{aligned}\text{Prob}(X_t = m) &= \left(\sum_{n_1+\cdots+n_{t-1}=n-m} \frac{(n-m)!}{n_1!\cdots n_{t-1}!} p_1^{n_1}\cdots p_{t-1}^{n_{t-1}}\right) \frac{n!}{m!(n-m)!} p_t^m \\ &= (p_1+\cdots+p_{t-1})^{n-m} \cdot \binom{n}{m} p_1^m;\end{aligned}$$

however, as $p_1+p_2+\cdots+p_t = 1$, we have $p_1+\cdots+p_{t-1} = 1-p_t$, and thus we finally obtain the solution

$$\text{Prob}(X_t = m) \ = \ \binom{n}{m} p_t^m (1-p_t)^{n-m}.$$

There is a more elegant way to see this without resorting to all the computations above. A multinomial with t probabilities $p_1,\dots,p_t$ models outcomes with t possibilities; for example, we might have t candidates and these are their support levels (or perhaps we have a strange die and these are the probabilities of a face landing up). When we sum all variables but one, we go from having t options to two options (either t or not t); it shouldn't be a surprise that this collapses to a binomial, as we are now lumping together all opposition.

Section 3.6: Problem 7: We are given that the joint mass function of X and Y is

$$f_{X,Y}(x,y) \ = \ \log_{10}\left(1+\frac{1}{10x+y}\right)$$

for $x \in \{1,\dots,9\}$ and $y \in \{0,\dots,9\}$. As a nice exercise, one should sum this and make sure it is a mass function. To find the marginal of X we sum over all Y; in other words, we want the probability $X = x$ and the value of Y is immaterial. Thus

$$f_X(x) \ = \ \sum_{y=0}^{9} \log_{10}\left(\frac{10x+y+1}{10x+y}\right).$$

There are two natural ways to do this sum. The first is to use $\log_{10}(A/B) = \log_{10} A - \log_{10} B$ and notice that we have a telescoping sum; the second is to note that the sum of logarithms is the logarithm of the product. In the latter approach, we find

$$f_X(x) \ = \ \log\left(\prod_{y=0}^{9} \frac{10x+y+1}{10x+y}\right).$$

The product is

$$\frac{10x+1}{10x} \cdot \frac{10x+2}{10x+1} \cdots \frac{10x+10}{10x+9};$$

the products cancel in pairs and all that remains is $\frac{10x+10}{10x} = \frac{x+1}{x}$. Thus the mass function is

$$f_X(x) \ = \ \log_{10}\left(\frac{1+x}{x}\right)$$

if $x \in \{1, \dots, 9\}$ and 0 otherwise. (A similar calculation shows that the sum of this over x equals 1, and thus our proposed function is indeed a probability mass function.)

The mean is just

$$\sum_{x=1}^{9} x \log_{10}\left(\frac{x+1}{x}\right) \ \approx \ 3.44024.$$

Can we somehow approximate this? Our sum is just

$$\begin{aligned}
&\log_{10}(2) - \log_{10}(1) \\
&2\log_{10}(3) - 2\log_{10}(2) \\
&3\log_{10}(4) - 3\log_{10}(3) \\
\vdots& \\
&9\log_{10}(10) - 9\log_{10}(9)\,.
\end{aligned}$$

Note this simplifies; instead of everything in the middle canceling we just get each once, and the mean is

$$9\log_{10}(10) - \sum_{k=1}^{9}\log_{10} k \ = \ \log_{10}(10^9) - \log_{10}\left(\prod_{k=1}^{9} k\right) \ = \ \log_{10}\frac{10^9}{9!} \ \approx \ 3.44024.$$

It is interesting to compare this answer to the average mantissa of a system satisfying Benford's law. Remember we may write any positive number x as $x = M_{10}(x)10^k$, where $M_{10}(x)$ is the mantissa of x (and lives in $[1,10)$) and k is an integer. For example, 1701.24601 = 1.70124601 $\cdot 10^3$. The density function for the mantissa is frequently $\frac{1}{x\log 10}$. Thus the expected value of the mantissa is

$$\int_1^1 0x \cdot \frac{1}{x\log 10}dx \ = \ \frac{9}{\log 10} \ \approx \ 3.90865,$$

which not surprising is a bit higher than what we calculated before. The reason it is higher is that if we only care about the first digit, then a number like 1.9997 counts as a first digit of 1, even though it is quite close to 2.

7. HW #7

Homework: Due Thursday October 29 (though you may place in my mailbox anytime up till 10am on Friday 10/30): Section 4.7: #2. Section 3.11: #14. Section 4.14: #35, #45bc.

Section 4.7: Problem 2: We are given that X and Y are independent exponential random variables with parameter 1; thus their joint density is

$$f_{X,Y}(x,y) \ = \ \begin{cases} e^{-x}e^{-y} & \text{if } x,y \geq 0 \\ 0 & \text{othwerwise.} \end{cases}$$

We now set $U = X+Y$ and $V = \frac{X}{X+Y}$; note that $0 \leq V \leq 1$ as $0 \leq X,Y$. To find the joint density of U and V, we need the Jacobian of the change of variables. We are given

$$T(X,Y) \ = \ (U(X,Y), V(X,Y)) \ = \ \left(X+Y, \frac{X}{X+Y}\right);$$

we need to invert this relation and solve for X and Y in terms of U and V. As $U = X+Y$, we may rewrite $V = \frac{X}{X+Y}$ as $V = X/U$, which means $X = UV$. Now that we know X in terms of U and V, we substitute into $U = X+Y$ to find $U = UV+Y$, or $Y = U - UV$. Thus

$$T^{-1}(U,V) \ = \ (X(U,V), Y(U,V)) \ = \ (UV, U-UV).$$

We can now calculate the Jacobian J, which tells us how the volume element transforms (explicitly, $dxdy = |J|dudv$). We have

$$J \ = \ \left| \begin{array}{cc} \frac{\partial X}{\partial U} & \frac{\partial Y}{\partial V} \\ \frac{\partial X}{\partial U} & \frac{\partial Y}{\partial V} \end{array} \right| \ = \ \left| \begin{array}{cc} V & U \\ 1-V & -U \end{array} \right| \ = \ -UV - U(1-V) \ = \ -U.$$

Thus the joint density of U and V is

$$f_{U,V}(u,v) \ = \ \begin{cases} f_{X,Y}(X(U,V),Y(U,V)) \cdot U & \text{if } U \geq 0 \text{ and } 0 \leq V \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

To find the marginal of V we integrate out U. If $v \notin [0,1]$ the answer is zero, and for $v \in [0,1]$ we have

$$\begin{aligned} f_V(v) &= \int_{u=0}^{\infty} f_{U,V}(u,v)du \\ &= \int_{u=0}^{\infty} e^{-x(u,v)}e^{-y(u,v)} \cdot udu \\ &= \int_{u=0}^{\infty} e^{-uv}e^{-(u-uv)}udu \\ &= \int_{u=0}^{\infty} e^{-u}udu \ = \ 1. \end{aligned}$$

There are many ways to see the last integral is 1. We can integrate by parts, we can note it is the mean of the standard exponential (i.e., the exponential with $\lambda = 1$), or we could observe that it is $\Gamma(1)$ which is $0! = 1$. We have thus shown that

$$f_V(v) \ = \ \begin{cases} 1 & \text{if } 0 \leq v \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

which proves that V is uniformly distributed on $[0,1]$. One interesting application is that if, somehow, we could generate independent values from the standard exponential, we could combine those to get a uniformly distributed random variable.

Section 3.11: Problem 14: Let $X_1, \dots, X_n$ be independent Bernoulli random variables, where $X_k \sim \mathrm{Bern}(p_k)$. By linearity of expectation, if $Y = X_1 + \cdots + X_n$ then we have

$$\mathbb{E}[Y] \ = \ \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \ = \ p_1 + \cdots + p_n,$$

as $\mathbb{E}[X_k] = p_k$. To see the later, recall the definitions: $X_k = 1$ with probability p_k and 0 with probability $1 - p_k$, and thus $\mathbb{E}[X_k] = 1 \cdot p_k + 0 \cdot (1 - p_k) = p_k$.

To compute the variance, we use the variance of a sum of independent random variables is the sum of the random variables. As $\mathrm{Var}(X_k) = p_k(1 - p_k)$, we find

$$\mathrm{Var}(Y) \ = \ \sum_{k=1}^{n} p_k(1 - p_k).$$

For a given mean of Y, what choices of p_k correspond to the largest possible variance? We first claim that there must be at least one choice which gives a maximum variance. To see this, we appeal to a result from real analysis: a continuous function on a compact set (i.e., a set that is closed and bounded) attains its maximum and minimum values.

It turns out to be sufficient to study the special case when $n = 2$; before explaining why, we'll analyze this case in detail. We give the 'standard' proof using techniques from calculus. While the idea is simple, the algebra quickly gets involved and tedious, though everything does work out if we're patient enough. As this much algebra is unenlightening, we give an alternate, simpler proof below as well.

First proof: long algebra. We first give the standard proof that one might give after taking a calculus class. Namely, we convert everything to a function of one variable, and just plow ahead with the differentiation, finding the critical points and comparing the values at the critical points to the end-points. While this is exactly what we've been taught to do in calculus, we'll quickly see the algebra becomes involved and unenlightening, and thus we will give *many* alternate proofs afterwards!

Our situation is that we have $p_1 + p_2 = \mu$ and we want to maximize $p_1(1 - p_1) + p_2(1 - p_2)$. As $p_2 = \mu - p_1$, we must maximize

$$\begin{aligned} g(p_1) &= p_1(1 - p_1) + (\mu - p_1)(1 - \mu + p_1) \\ &= p_1 - p_1^2 + \mu(1 - \mu) - p_1(1 - \mu) + p_1\mu - p_1^2 \\ &= 2p_1\mu - 2p_1^2 + \mu(1 - \mu). \end{aligned}$$

To find the maximum, calculus tells us to find the critical points (the values of p_1 where $g'(p_1) = 0$) and compare that value to the endpoints (which for this problem would be $p_1 = \max(0, \mu - 1)$ and $p_1 = \min(\mu, 1)$). We have $g'(p_1) = 2\mu - 4p_1$, so the critical point is $p_1 = \mu/2$ which gives $g(\mu/2) = \mu - \frac{\mu^2}{2}$. Straightforward algebra now shows that this is larger than the boundary values. As $g(p_1) = g(1 - p_1)$, it suffices to check the lower bounds. If $p_1 = 0$ that means $0 \le \mu \le 1$, and in this case $p_2 = \mu$ so $g(0) = \mu(1 - \mu) = \mu - \mu^2$, which is clearly smaller than $g(\mu/2) = \mu - \frac{\mu^2}{2}$. Similarly if $p_1 = \mu - 1$ (which implies $1 \le \mu \le 2$) then $p_2 = 1$ and thus $g(\mu - 1) =$

$(\mu-1)(2-\mu)+0=-\mu^2+3\mu-2$. If this were larger than $g(\mu/2$, we would have the following chain:

$$\begin{array}{rcl} -\mu^2+3\mu-2 & > & \mu-\frac{\mu^2}{2} \\ 0 & > & \frac{\mu^2}{2}-2\mu+2 \\ 0 & > & \mu^2-4\mu+4 \\ 0 & > & (\mu-2)^2, \end{array}$$

which is impossible. Thus, after tedious but straightforward algebra, we see the maximum value occurs not at a boundary point but at the critical point $p_1 = \mu/2$, which implies $p_2 = \mu/2$ as well.

We now consider the case of general n. Imagine we are at the maximum variance with values $\mathfrak{p}_1, \cdots, \mathfrak{p}_n$. If any two of the $\mathfrak{p}_k$'s were unequal (say the i and j values), by the argument above (in the case of just two values) we could increase the variance by replacing $\mathfrak{p}_i$ and $\mathfrak{p}_j$ with $\frac{\mathfrak{p}_i+\mathfrak{p}_j}{2}$. Thus the maximum value of the variance occurs when all are equal.

Second proof: cleaner algebra. As the algebra is a bit tedious, we give another approach. Imagine (back in the $n=2$ case) that $p_1 \neq p_2$. Let's write $p_1 = \frac{\mu}{2}+x$ and $p_2 = \frac{\mu}{2}-x$. We need to show the variance is maximized when $x=0$. If $x=0$ the variance is just $\mu-\frac{\mu^2}{2}$, while for general x it is

$$\left(\frac{\mu}{2}+x\right)\left(1-\frac{\mu}{2}-x\right)+\left(\frac{\mu}{2}-x\right)\left(1-\frac{\mu}{2}+x\right) \ = \ \mu-\frac{\mu^2}{2}-2x^2,$$

where the last step follows from multiplying everything out. Thus the variance is maximized in this case when $x=0$. Note how much faster this approach is. We included the first approach as this is what we're taught in calculus, namely find the critical points and check the boundary points; however, especially in instances where we have some intuition as to what the answer should be, there are frequently better ways of arranging the algebra.

Third proof: Lagrange multipliers. We give one more proof, though here the prerequisites are more. We use Lagrange multipliers: we want to maximize $f(p_1,p_2) = p_1(1-p_1)+p_2(1-p_2)$ subject to $g(p_1,p_2) = p_1+p_2-\mu = 0$. We need $\nabla f = \nabla g$, so

$$\begin{array}{rcl} f(p_1,p_2) & = & p_1-p_1^2+p_2-p_2^2 \\ g(p_1,p_2) & = & p_1+p_2-\mu \\ \nabla f(p_1,p_2) & = & (1-2p_1, 1-2p_2) \\ \nabla g(p_1,p_2) & = & (1,1). \end{array}$$

As $\nabla f = \lambda g$ and $\nabla g(p_1,p_2) = (1,1)$, we find $1-2p_1 = 1-2p_2$ or $p_1 = p_2$ as claimed. Note how readily this generalizes to n variables, as in this case we would have

$$\begin{array}{rcl} \nabla f(p_1,\dots,p_n) & = & (1-2p_1,\dots,1-2p_n) \\ \nabla g(p_1,\dots,p_n) & = & (1,\dots,1), \end{array}$$

which implies all the p_i's are equal.

Fourth proof: geometry. We give yet another proof in the case $n = 2$ and $p_1 + p_2 = \mu$. We are trying to maximize

$$p_1(1-p_1) + p_2(1-p_2) \ = \ p_1 - p_1^2 + p_2 - p_2^2 \ = \ \mu - (p_1^2 + p_2^2).$$

As we are subtracting $p_1^2 + p_2^2$, we want that to be as small as possible. We may interpret this as the distance of the point (p_1, p_2) from the origin, given that $p_1 + p_2 = \mu$. Geometrically it should be clear that the closest point to the origin is the midpoint of the line from $(0, \mu)$ to $(\mu, 0)$; if not and if we need to resort to calculus, this is at least an easier problem. Namely, let $p_2 = \mu - p_1$ so we are trying to minimize

$$\mu - (p_1^2 + (\mu - p_1)^2) \ = \ \mu - \mu^2 - (2p_1^2 - 2\mu p_1) \ = \ \mu - \mu^2 - 2p_1(p_1 - \mu).$$

We thus need to minimize the value of the quadratic $p_1(p_1 - \mu)$; as the roots of this are 0 and μ, the minimum is at the vertex which is at the midpoint of the roots, namely $p_1 = \mu/2$. In general, we are trying to minimize the function $\mu - (p_1^2 + \cdots + p_n^2)$ subject to $0 \le p_1, \dots, p_n \le 1$ and $p_1 + \cdots + p_n = \mu$. This is equivalent to finding the point on the hyperplane closest to the origin in n-dimensional space, which is given by the point where they are all equal.

Finally, is this result surprising? If ever a $p_k = 0$ or 1, then there would be no variation in the contribution from X_k. Thus the variance will be smallest when all the p_k's are in $\{0, 1\}$.

Section 4.14: Problem 35: The marriage or secretary problems is one of the more famous probability exercises. Though the terminology changes based on who is presenting it, the basic idea is as follows. We have a known number, say n, of objects (which are frequently candidates for a job, suitors, or preferences). We can always rank and order any collection of these, and there are no ties. For example, if the candidates are Alice, Bob, Charlie, Ethelbert and Daneel, our ranking may be Bob, Ethelbert, Daneel, Alice and Daneel; this means we prefer Bob over all, but we do not say by how much we prefer Bob to Ethelbert.

We are now shown the objects one at a time. The goal is to design a strategy so that we stop at the best alternative. Unfortunately for us, we are forced to make an accept / reject decision on each candidate the moment we see them. Thus, if the first person we see is Ethelbert, we must then and there choose whether or not to keep Ethelbert, or take some future unspecified candidate. This is why this is often called the marriage problem (once you reject a suitor, it is unlikely they will look favorably on you again).

We desire a strategy that maximizes the chance of ending up with the best candidate. It would be so easy if we could just see all the candidates and then decide; sadly, we must make our decision on each candidate immediately upon seeing them. One strategy is to just always take the first person (or always take the fifth, or eighteenth, et cetera). This will give us the best candidate with probability $1/n$, which is not too impressive for n large. Can we do better?

The following strategy, called S_k, is frequently used. Let us look at the first k people, and then we'll choose the first person we see from this point onward who is better than the best we've seen in the first k. How good is this strategy? Clearly it is bad when the best person is one of the first k, as then we'll never take them. This happens with probability k/n, and so k/n of the time we always lose.

If the best person is in position $k+1$, however, we always win. In general, the best person will be in position m. We have already analyzed how good our strategy is when m happens to be in $\{1, 2, \dots, k, k+1\}$; what about other m? Assume $m \geq k+1$ is the location of the best person. Our strategy S_k results in our selecting the best person if and only if there is no one among people $k+1, \dots, m-1$ who is better than the best person in the first k. For example, imagine there are 100 candidates and we interview the first ten. Let's say Julia was the best we saw in the first 10. The best candidate overall happens to be Zeke, who is in position 27. Thus we will only end up choosing Zeke if Julia is better than all the candidates from the 11th to the 26th position, as right now we are searching for the first person better than Julia. What is the probability that the best person among the first 26 candidates happens to lie in the first 10? The answer is just 10/26.

In general, if the best person is at position m then we select the best person precisely when the best person among the first $m-1$ is in the first k people. The probability the best of the first $m-1$ is in the first k is just $\frac{k}{m-1}$. We therefore find that the probability strategy S_k wins is

$$\begin{aligned}
\text{Prob}(S_k \text{ wins}) &= \sum_{m=k+1}^{n} \text{Prob}(\text{win}|\text{best at } m) \cdot \text{Prob}(\text{best at } m) \\
&= \sum_{m=k+1}^{n} \frac{k}{m-1} \cdot \frac{1}{n} \\
&= \frac{k}{n} \sum_{m=k+1}^{n} \frac{1}{m-1} \\
&= \frac{k}{n} \left(\sum_{m=1}^{n-1} \frac{1}{m} - \sum_{m=1}^{k-1} \frac{1}{m} \right) \\
&= \frac{k}{n} \left(H_{n-1} - H_{k-1} \right),
\end{aligned}$$

where

$$H_\ell = \sum_{m=1}^{\ell} \frac{1}{m}$$

is the ℓ^{th} harmonic number, which is approximately $\log \ell$ for ℓ large. Thus

$$\text{Prob}(S_k \text{ wins}) \approx \frac{k}{n} \log\left(\frac{n-1}{k-1}\right) = \frac{\log\left(\frac{n-1}{k-1}\right)}{\frac{n}{k}}.$$

For n and k large, we may replace $n-1$ with n and $k-1$ with k. Thus we are trying to optimize $g(x) = \frac{\log x}{x}$, where $1 \leq x = \frac{n}{k} \leq n$. To find where a function is largest, we check the critical and endpoints. Letting $g(x) = \frac{\log x}{x}$, we see the endpoints give

$g(1) = 0$, $g(n) = \frac{\log n}{n}$. As

$$g'(x) \ = \ \frac{\frac{1}{x} \cdot x - \log x \cdot 1}{x^2},$$

$g'(x) \ = \ 0$ implies $\log x \ = \ 1$ or $x \ = \ e$. Thus the optimal k is about n/e, and the probability we end up with the best is approximately

$$\frac{\log\left(\frac{n}{n/e}\right)}{\frac{n}{n/e}} \ = \ \frac{\log e}{e} \ = \ \frac{1}{e} \ \approx \ 36.8\%.$$

This is amazing. The naive strategy of always taking a fixed position (such as always take the first candidate) gets the best $1/n$ of the time. If we look at the first $1/e$ percent and then take the first one better than the best here, we end with the best approximately $1/e$ percent of the time!

Advanced note. We have to be a little careful, as k must be an integer. Though we have made some approximations, we see the derivative of the probability of winning is $(1 - \log x)/x^2$, with $x = n/k$. We see the derivative is positive for $x < e$ and negative for $x > e$. Thus the plot looks like an inverted u, and thus the integer maximum is either the integer immediately to the right or left of the critical point.

There are lots of generalizations. We discuss in detail one below, and leave the others for the reader to explore.

Getting one of the top two. The next question would be: what strategy gives the largest probability that we end up with either the best or second best candidate? The answer turns out to be over 50%! We assume again we have a simple strategy of interviewing the first k candidates, and afterwards discuss some variants. We'll denote the location of the best and second best candidates as m_1 and m_2. We analyze the problem in greater detail then needed to get a sense of the answer.

- If both $m_1, m_2 \le k$ we always lose, and this happens with probability $\binom{k}{2}/\binom{n}{2} = \frac{k(k-1)}{n(n-1)}$. We can see this in two ways. The first is there are $\binom{n}{2}$ ways to choose where to put two people, and $\binom{k}{2}$ ways to put two people in two of the first k positions. Alternatively, the probability the first person is in the first k is $\frac{k}{n}$, and then the probability that the second person is also in the first k is $\frac{k-1}{n-1}$ (as one slot has been filled. If k again is of the same order of magnitude as n, then this is a significant probability of failure.
- If the best is in the first k and the second is not, we lose unless the second best happens to be in the final position. Thus the probability we win in this case is $\frac{k}{n}\frac{1}{n}$.
- If the second best candidate is in the first k and the best is not, we automatically win with this strategy! The probability of this happening is $\frac{k}{n}\frac{n-k}{n} = \frac{k}{n}\left(1 - \frac{k}{n}\right)$. If k is of the same order of magnitude as n, then this will be a significant probability of success.

- Finally, we are reduced to analyzing the case when the top two candidates are not in the first k. The probability of success in this case is

$$\sum_{m_1=k+1}^{n-1} \sum_{m_2=m_1+1}^{n} \text{Prob}(\text{win}|\{\text{best, second}\} = \{m_1, m_2\}) \cdot \text{Prob}(\{\text{best, second}\} = \{m_1, m_2\})$$

The probability that the best and second best are in positions $\{m_1, m_\}$ is just $\frac{2}{n}\frac{1}{n-1}$ (there are two positions where we may place the best candidate among the n people, and then one position remaining for the second best candidate; alternatively, we could view this as $1/\binom{n}{2}$). What is the probability we win, given that the best two candidates are at positions $\{m_1, m_2\}$? The argument is the same as before – we need the best person among the first $m_1 - 1$ candidates to be in the first k candidates. Thus, the probability we win in this case is just $\frac{k}{m_1-1}$, and so summing over m_1 and m_2 we find

$$\begin{aligned} \sum_{m_1=k+1}^{n-1} \sum_{m_2=m_1+1}^{n} \frac{k}{m_1-1}\frac{2}{n(n-1)} &= \frac{2k}{n(n-1)} \sum_{m_1=k+1}^{n-1} \frac{1}{m_1-1} \sum_{m_2=m_1+1}^{n} 1 \\ &= \frac{2k}{n(n-1)} \sum_{m_1=k+1}^{n-1} \frac{1}{m_1-1}(n-m_1). \end{aligned}$$

We do one of the most common, useful tricks to evaluate the sum – we write $n - m_1$ as $n - 1 - (m_1 - 1)$. The reason we do this is that the denominator is $m_1 - 1$, and this will lead to nice simplifications. We thus find the probability of winning, in this case, is

$$\begin{aligned} & \frac{2k}{n(n-1)} \left[(n-1) \sum_{m_1=k+1}^{n-1} \frac{1}{m_1-1} - \sum_{m_1=k+1}^{n-1} 1\right] \\ = \; & \frac{2k}{n(n-1)} \left[(n-1)(H_{n-2} - H_{k-1}) - (n-1-k)\right] \\ \approx \; & \frac{2k}{n} \left[\log\frac{n-2}{k-1} - \left(1 - \frac{k}{n-1}\right)\right] \\ \approx \; & \frac{2k}{n} \left[\log\frac{n}{k} - 1 + \frac{k}{n}\right]. \end{aligned}$$

Combining all the different probabilities, we see the probability of winning is

$$\text{Prob}(\text{win}) \;\approx\; \frac{k}{n^2} + \frac{k}{n}\left(1 - \frac{k}{n}\right) + \frac{2k}{n}\left[\log\frac{n}{k} - 1 + \frac{k}{n}\right].$$

As k will be of the same size as n, the k/n^2 term is negligible and may safely be ignored. If we let $x = n/k$ as before, we see we must optimize the function

$$g(x) \;=\; \frac{1}{x}\left(1 - \frac{1}{x}\right) + \frac{2}{x}\left(\log x - 1 + \frac{1}{x}\right), \quad 1 \le x \le n.$$

The algebra and calculus is easier if instead we let $y = k/n = 1/x$, as this gives

$$h(y) \ = \ y(1-y) + 2y\left(-\log y - 1 + y\right), \quad \frac{1}{n} \le y \le 1.$$

After some algebra, we see the derivatives are

$$h'(y) \ = \ -3 + 2y - 2\log y, \quad h''(y) \ = \ 2 - \frac{2}{y}.$$

Numerically solving gives $y \approx 0.30171$, and we easily see this is a maximum. Further, this is clearly better than the endpoint strategies of $y = 1/n$ or $y = 1$, and thus the maximum probability is when $y \approx 0.30171$. Substituting this into our formula, we find the probability of winning with this strategy is about 0.51239, or in other words we have greater than a 50% chance of getting one of the top two candidates!

Let's summarize our results:

Goal	k/n (i.e, percent look at)	Probability of winning
Best candidate	About $1/e \approx 36.8\%$	About 36.8%
One of top two	About 30.2%	About 51.2%

What if we applied our original strategy of looking at the first $k \approx n/e$ people – what would be the probability that we end up with one of the two best? Substituting in $y = 1/e$ gives a probability of $\frac{1+e}{e^2} = \frac{1}{e} + \frac{1}{e^2}$. As in this strategy we have a probability of $1/e$ of ending up with the best person, we must therefore have a probability of $1/e^2$ of ending up with the second best. A natural question is what do you expect the probability to be of ending up with one of the best ℓ people given that we look at the first n/e people?

It is also interesting to note that the difference in the probability of getting one of the top two if we look at $n/e \approx .368n$ versus looking at $.302n$ is small, namely about 50.3% to 51.2%.

In general, if we want to get one of the ℓ best, about how many people should we interview? If we want one of the ℓ best, would it perhaps be better to interview k people and then take the *second* (or maybe even the third, the fourth, ...) person better than the best we've seen?

Another generalization is in determining the probability that we end with a candidate in the top ϵ percent. How many people would we interview in this case? What would our probability of success be?

Finally, we can consider the generalization where we have a quantifiable ranking *and* knowledge that the candidates are drawn from a fixed distribution, for example the uniform distribution on $[a, b]$, though we do not know a or b. We then interview the first k people and try to estimate the values of a and b. This method involves order statistics, which appear in problems ranging from the distribution of sample medians to inferences in statistics.

Section 4.14: Problem 45bc: The kurtosis of a random variable X is defined by $\mathrm{kur}(X) := \mathbb{E}[(X-\mu)^4]/\sigma^4$, where μ is the mean and σ is the standard deviation. The kurtosis measures how much probability we have in the tails.

Let $X \sim \mathrm{Poiss}(\lambda)$, so the mass function is $f(n) = \lambda^n e^{-\lambda}/n!$ for $n \geq 0$ and 0 otherwise. For a Poisson random variable with parameter λ, the mean is λ and the standard deviation is $\sqrt{\lambda}$ (or equivalently the variance is λ), and thus

$$\mathrm{kur}(X) \ = \ \frac{\sum_{n=0}^\infty (n-\lambda)^4 \lambda^n e^{-\lambda}/n!}{\lambda^2}.$$

There are several ways to try and analyze this. One way is to expand out $(n-\lambda)^4$. Whenever we have an n, we can cancel that with the n in $n!$, and we are left with terms such as $n^k\lambda^j/(n-1)!$. We could then write n as $(n-1)+1$, expand and do some more canceling. While this will work, the algebra becomes tedious. The point of this exercise is to see that, while there are numerous ways to solve a problem, it is important to weigh their advantages and disadvantages. For instance, we can either make the linear combinations easy at the cost of more involved differentiation, or we can have easier combinations at the expense of more tedious differentiation. For this problem, it seems as if the easiest algebra is when we make the differentiation hard but the combinations easy. It takes awhile to develop a feel for which approach will be most tractable for a given problem. This is one reason why we provide so many different solutions.

First solution. One of the best ways to compute the moments of Poisson (and other discrete) random variables is through differentiating identities. Consider the identity

$$e^x \ = \ \sum_{n=0}^\infty \frac{x^n}{n!}.$$

We could keep applying the operator $x\frac{d}{dx}$ to this and obtain the moments, and then by expanding $(n-\lambda)^4$ piece everything together. A faster way is to apply the operator $-\lambda + x\frac{d}{dx}$ four times and then set $x=\lambda$. If we do that we obtain

$$\left(-\lambda + x\frac{d}{dx}\right)\left(-\lambda + x\frac{d}{dx}\right)\left(-\lambda + x\frac{d}{dx}\right)\left(-\lambda + x\frac{d}{dx}\right)e^x\Big|_{x=\lambda} \ = \ \sum_{n=0}^\infty (n-\lambda)^4 \cdot \frac{\lambda^n}{n!}.$$

After some long but standard differentiation, we find the derivative above equals

$$e^x\left(\lambda^4 - 4\lambda^3 x + 6\lambda^2 x(1+x) - 4\lambda x(1+3x+x^2) + x(1+7x+6x^2+x^3)\right);$$

setting $x=\lambda$ gives

$$\lambda e^\lambda + 3\lambda^2 e^\lambda \ = \ \sum_{n=0}^\infty (n-\lambda)^4 \cdot \frac{\lambda^n}{n!},$$

which means the kurtosis is

$$\mathrm{kur}(X) \ = \ \frac{e^{-\lambda}}{\lambda^2}\left(\lambda e^\lambda + 3\lambda^2 e^\lambda\right) \ = \ 3 + \frac{1}{\lambda}.$$

Second solution. In terms of keeping the algebra simple, it might be easier to expand $(n-\lambda)^4$ and apply the operator $x\frac{d}{dx}$ four times.

Third solution. Another possibility is to apply d/dx four times and then build back. For example, we start with

$$e^x \ = \ \sum_{n=0}^\infty \frac{x^n}{n!}.$$

Differentiating with respect to x once gives

$$e^x \ = \ \sum_{n=0}^\infty n \cdot \frac{x^{n-1}}{n!}.$$

Taking $x = \lambda$ and multiplying both sides by $\lambda e^{-\lambda}$ gives

$$\lambda e^{-\lambda} \cdot e^\lambda \ = \ \sum_{n=0}^\infty n \cdot \frac{\lambda^n e^{-\lambda}}{n!} \ = \ \mathbb{E}[X],$$

which implies the mean is λ. If we differentiate e^x twice with respect to x, we find

$$e^x \ = \ \sum_{n=0}^\infty n(n-1) \cdot \frac{x^{n-2}}{n!} \ = \ \sum_{n=0}^\infty n^2 \cdot \frac{x^{n-2}}{n!} - \sum_{n=0}^\infty n \cdot \frac{x^{n-2}}{n!}.$$

Taking $x = \lambda$ again and multiplying both sides by $\lambda e^{-\lambda}$ gives

$$\lambda^2 e^{-\lambda} e^\lambda \ = \ \sum_{n=0}^\infty n^2 \cdot \frac{\lambda^n e^{-\lambda}}{n!} - \sum_{n=0}^\infty n \cdot \frac{\lambda^n e^{-\lambda}}{n!};$$

as the last sum is λ, we find

$$\mathbb{E}[X^2] \ = \ \sum_{n=0}^\infty n^2 \cdot \frac{\lambda^n e^{-\lambda}}{n!} \ = \ \lambda^2 + \lambda.$$

Continuing in this way we can get $\mathbb{E}[X^3]$ and $\mathbb{E}[X^4]$, and then substitute into

$$\mathbb{E}[(X-\mu)^4] \ = \ \mathbb{E}[X^4] - 4\mu\mathbb{E}[X^3] + 6\mu^2\mathbb{E}[X^2] - 4\mu^3\mathbb{E}[X] + \mu^4.$$

Fourth solution. For our fourth solution, we use some ideas from linear algebra. We start, as always, with the identity $e^x = \sum_{n=0}^{\infty} x^n/n!$, and we differentiate this 4 times:

$$\begin{aligned}
e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \\
e^x &= \sum_{n=0}^{\infty} n \cdot \frac{x^{n-1}}{n!} \\
e^x &= \sum_{n=0}^{\infty} n(n-1) \cdot \frac{x^{n-2}}{n!} \\
e^x &= \sum_{n=0}^{\infty} n(n-1)(n-2) \cdot \frac{x^{n-3}}{n!} \\
e^x &= \sum_{n=0}^{\infty} n(n-1)(n-2)(n-3) \cdot \frac{x^{n-4}}{n!}.
\end{aligned}$$

We take $x = \lambda$ and multiply the k^{th} equation above by λ^k, and find

$$\begin{aligned}
e^\lambda &= \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \\
\lambda e^\lambda &= \sum_{n=0}^{\infty} n \cdot \frac{\lambda^n}{n!} \\
\lambda^2 e^\lambda &= \sum_{n=0}^{\infty} (n^2 - n) \cdot \frac{\lambda^n}{n!} \\
\lambda^3 e^\lambda &= \sum_{n=0}^{\infty} (n^3 - 3n^2 + 2n) \cdot \frac{\lambda^n}{n!} \\
\lambda^4 e^\lambda &= \sum_{n=0}^{\infty} (n^4 - 6n^3 + 11n^2 - 6n) \cdot \frac{\lambda^n}{n!}.
\end{aligned}$$

We want to evaluate

$$\frac{e^{-\lambda}}{\lambda^2} \sum_{n=0}^{\infty} (n-\lambda)^4 \cdot \frac{\lambda^n}{n!} \ = \ \frac{e^{-\lambda}}{\lambda^2} \sum_{n=0}^{\infty} (n^4 - 4n^3\lambda + 6n^2\lambda^2 - 4n\lambda^3 + \lambda^4) \cdot \frac{\lambda^n}{n!}.$$

We write $n^4 - 4n^3\lambda + 6n^2\lambda^2 - 4n\lambda^3 + \lambda^4$ as a linear combination of the terms above. This is just solving a system of equations (for example, we may regard $n^4 - 4n^3\lambda + 6n^2\lambda^2 - 4n\lambda^3 + \lambda^4$ as the vector $(1, -4, 6, -4, 1, 0)$, with the last component 0 as there is no constant term). Solving the associated system of equations gives

$$n^4 - 4n^3\lambda + 6n^2\lambda^2 - 4n\lambda^3 + \lambda^4$$

equals

$$\begin{aligned}
& 1\cdot(n^4-6n^3+11n^2-6n)\\
&+(6-4\lambda)\cdot(n^3-3n^2+2n)\\
&+(7-12\lambda+6\lambda^2)\cdot(n^2-n)\\
&+(1-4\lambda+6\lambda^2-4\lambda^3)\cdot n\\
&+a^4\cdot 1
\end{aligned}$$

and thus the kurtosis is

$$\begin{aligned}
& \frac{e^{-\lambda}}{\lambda^2}\Big[1\cdot\lambda^4 e^{\lambda}+(6-4\lambda)\lambda^3 e^{\lambda}+(7-12\lambda+6\lambda^2)\lambda^2 e^{\lambda}+\\
&\qquad (1-4\lambda+6\lambda^2-4\lambda^3)\lambda e^{\lambda}+1e^{\lambda}\Big]\\
=\ & \frac{1}{\lambda^2}\left[3\lambda^2+\lambda\right]\ =\ 3+\frac{1}{\lambda}.
\end{aligned}$$

Consider $X\sim\mathrm{Exp}(\lambda)$, which has mean and standard deviation both equal to $1/\lambda$. The density of X is $f_X(x)=\lambda e^{-\lambda x}$ for $x\geq 0$ and 0 otherwise, and thus we have

$$\begin{aligned}
\mathrm{kur}(X) &= \frac{\int_0^\infty\left(x-\frac{1}{\lambda}\right)^4\lambda e^{-x\lambda}dx}{\frac{1}{\lambda^4}}\\
&= \int_0^\infty(\lambda x-1)^4 e^{-\lambda x}\lambda dx\\
&= \int_0^\infty(u-1)^4 e^{-u}du.
\end{aligned}$$

There are several ways to proceed at this point. We can integrate by parts or we can expand out. We choose to expand out, as we will recognize the answer. We find

$$\begin{aligned}
\mathrm{kur}(X) &= \int_0^1(u^4-4u^3+6u^2-4u+1)e^{-u}du\\
&= \int_0^1 u^4e^{-u}du-4\int_0^1 u^3e^{-u}du+6\int_0^1 u^2e^{-u}du-4\int_0^1 ue^{-u}du+\int_0^1 e^{-u}du\\
&= \Gamma(5)-4\Gamma(4)+6\Gamma(3)-4\Gamma(2)+\Gamma(1),
\end{aligned}$$

where we are using

$$\Gamma(s)\ =\ \int_0^\infty e^{-u}u^{s-1}du,\quad \mathfrak{Re}(s)>0.$$

As $\Gamma(n+1)=n!$ for n a positive integer, we have

$$\mathrm{kur}(X)\ =\ 4!-4\cdot 3!+6\cdot 2!-4\cdot 1!+0!\ =\ 9.$$

8. HW #8

Due Thursday November 5: Section 5.1: #1c. Section 1.8: #23 (can you say anything about this sum when n is large?). Section 4.9: #6. Additional Problems: (1) Let $X_1, \dots, X_n$ be independent random variables having the standard exponential distribution. Using convolutions, find the density for $X_1 + X_2$; more generally, find the density for $X_1 + \cdots + X_n$. (2) Find the Fourier transforms of f and g, where f is the density of the uniform distribution on $[0, 1]$ and g is the density of the uniform distribution on $[-1/2, 1/2]$. (3) Let $X_1, \dots, X_n$ be n independent standard normals. Using convolutions, show $X_1^2 + X_2^2$ has a chi-square distribution with 2 degrees of freedom, and more generally that $X_1^2 + \cdots + X_n^2$ has a chi-square distribution with n degrees of freedom.

Section 5.1: Problem 1c: We want to find the generating function for $f(m) = (1 - p)p^{|m|}/(1 + p)$ for $m \in \{\dots, -1, 0, 1, \dots\}$. The generating function is defined by $G(s) = \mathbb{E}[s^X]$, so in our case we have

$$\begin{aligned} G(x) &= \mathbb{E}\left[s^X\right] \\ &= \sum_{m=-\infty}^{\infty} s^m \cdot \frac{(1-p)p^{|m|}}{1+p} \\ &= \frac{1-p}{1+p} \sum_{m=-\infty}^{\infty} s^m p^{|m|} \\ &= \frac{1-p}{1+p} \left[\sum_{m=0}^{\infty} (sp)^m + \sum_{m=0}^{\infty} (p/s)^m - 1\right] \\ &= \frac{1-p}{1+p} \left[\frac{1}{1-sp} + \frac{1}{1-(p/s)} - 1\right]. \end{aligned}$$

So long as $|sp| < 1$ then the first sum converges, while the second sum converges if $|p/s| < 1$. Combining these, we see the generating function $G(s)$ is well-defined so long as $p < |s| < 1/p$. Unlike previous problems, instead of having convergence in a ball about the origin we now have convergence in an annular (or donut) region.

To calculate the mean and the variance, we use the following formulas:

$$\mathbb{E}[X] \ = \ G'_X(1), \quad \text{Var}(X) \ = \ G''_X(1) + G'_X(1) - G'_X(1)^2.$$

We now see the power of generating functions; we can differentiate $G_X(s)$ easily, and this is much better than evaluating sums. Fortunately $G_X(s)$ is defined at $s = 1$ (which, provided $p < 1$, is always inside the annular region). We have

$$\mathbb{E}[X] \ = \ G'_X(1) \ = \ 0.$$

While we could perform the algebra to compute $G'_X(1)$, there is no need if we only care about the mean. The reason is the probability distribution is symmetric about $m = 0$. It is worth recording that

$$G_X(s) \ = \ \frac{1-p}{1+p} \left[\frac{p}{(1-sp)^2} - \frac{p}{(1-(p/s))^2 s^2}\right].$$

To calculate the variance, the only additional information we need is $G_X''(1)$, which is

$$\begin{aligned} G''(1) &= \frac{1-p}{1+p}\left[\frac{2p^2}{(1-ps)^3} + \frac{2p^2}{(1-(p/s))^3 s^4} + \frac{2p}{(1-(p/s))^2 s^3}\right] \\ &= \frac{1-p}{1+p} \cdot \frac{2p(1+p)}{(1-p)^3} \\ &= \frac{2p}{(1-p)^2}. \end{aligned}$$

Additional problem (1). We are given that $X_1, \dots, X_n$ are independent standard exponential random variables. Thus the density function for each is $f(x) = e^{-x}$ for $x \geq 0$ and 0 otherwise. The density for $X_1 + X_2$ is simply the convolution of f with itself, or

$$\begin{aligned} f_{X_1+X_2}(x) &= (f * f)(x) \\ &= \int_{-\infty}^{\infty} f(t) f(x-t) dt. \end{aligned}$$

As f vanishes whenever it is evaluated at a negative number, the factor $f(t)$ restricts the integration to be from 0 to ∞. The second factor, $f(x-t)$, is zero unless $t \leq x$. Thus for $x \geq 0$ we have

$$\begin{aligned} f_{X_1+X_2}(x) &= \int_0^{\mathbf{x}} e^{-t} e^{-(x-t)} dt \\ &= \int_0^{\mathbf{x}} e^{-x} dt \\ &= x e^{-x}. \end{aligned}$$

As a quick check, we test to make sure xe^{-x} is a probability distribution. It is non-negative on $[0, \infty)$, and it does integrate to 1 (it is just $\Gamma(2) = 1!$, or alternatively we could just integrate by parts).

One of the most common mistakes made by probability students is to forget that the density $f_{X_i}(x)$ is e^{-x_i} only when $x_i \geq 0$; in other words, it is common to mistakenly use this as the definition for all x. This cannot be right; note that as $x_i \to -\infty$ the factor e^{-x_i} tends to infinity, and is not integrable. In summary, a common pitfall is to say that

$$f_{X_1+X_2}(x) = \int_0^{\infty} e^{-t} \cdot e^{-(x-t)} dt = \int_0^{\infty} e^{-x} dt = e^{-x} \int_0^{\infty} dt,$$

and clearly there is no way this integral will be finite!

Let's calculate the density for $X_1 + X_2 + X_3$. As convolution is associative, the density is just $(f * f) * f$, so arguing as above we find

$$\begin{aligned} f_{X_1+X_2+X_3}(x) &= \int_{-\infty}^{\infty} f_{X_1+X_2}(t) f(x-t) dt \\ &= \int_0^x te^{-t} e^{-(x-t)} dt \\ &= \int_0^x te^{-x} dt \\ &= e^{-x} \frac{x^2}{2}. \end{aligned}$$

Let's do one more to make the pattern clear, and then we'll generalize our observations and prove them by induction. For $X_1 + \cdots + X_4$ we have the density is $(f*f*f)*f$, and thus

$$\begin{aligned} f_{X_1+\cdots+X_4}(x) &= \int_{-\infty}^{\infty} f_{X_1+X_2+X_3}(t) f(x-t) dt \\ &= \int_0^x \frac{t^2}{2}\, e^{-t} e^{-(x-t)} dt \\ &= e^{-x} \int_0^x \frac{t^2}{2} \\ &= e^{-x} \frac{x^3}{3!}. \end{aligned}$$

Based on the above calculations, we conjecture that the sum of n independent standard exponentials has density function $e^{-x} x^n / n!$. We now prove this by induction. We have done the basis case above. Assuming it holds for n, we must show it holds for $n + 1$. But

$$\begin{aligned} f_{X_1+\cdots+X_{n+1}}(x) &= \int_{-\infty}^{\infty} f_{X_1+\cdots+X_n}(t) f(x-t) dt \\ &= \int_0^x \frac{t^n}{n!} e^{-t} e^{-(x-t)} dt \\ &= e^{-x} \int_0^x \frac{t^n}{n!} \\ &= e^{-x} \frac{x^{n+1}}{n!}. \end{aligned}$$

This is the Gamma distribution, and is a famous, important density.

Additional problem (2). We calculate the Fourier transform of the uniform density on $[0, 1]$. We have

$$\widehat{f}(\xi) \;=\; \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx \;=\; \int_0^1 e^{-2\pi i x \xi} dx.$$

If $\xi = 0$ then the answer is clearly just 1. For other ξ, we use the fact that $e^{-i\theta} = \cos\theta - i\sin\theta$. For us, we have

$$e^{-2\pi i x\xi} \ = \ \cos(2\pi x\xi) - i\sin(2\pi x\xi),$$

and hence

$$\begin{aligned}\widehat{f}(\xi) &= \int_0^1 \left[\cos(2\pi x\xi) - i\sin(2\pi x\xi)\right] dx \\ &= \int_0^1 \cos(2\pi x\xi)dx - i\int_0^1 \sin(2\pi x\xi)dx \\ &= \left.\frac{\sin(2\pi x\xi)}{2\pi\xi}\right|_0^1 + i\left.\frac{\cos(2\pi x\xi)}{2\pi\xi}\right|_0^1 \\ &= \frac{\sin(2\pi\xi)}{2\pi\xi} + i\left(\frac{\cos(2\pi\xi)}{2\pi\xi} - \frac{1}{2\pi\xi}\right).\end{aligned}$$

Consider now the uniform distribution on $[-1/2, 1/2]$. The only thing that changes in the above analysis is the last step, where now instead of evaluating the integrals at 0 and 1 we evaluate at $-1/2$ and $1/2$. We thus find

$$\begin{aligned}\widehat{g}(\xi) &= \left.\frac{\sin(2\pi x\xi)}{2\pi\xi}\right|_{-1/2}^{1/2} + i\left.\frac{\cos(2\pi x\xi)}{2\pi\xi}\right|_{-1/2}^{1/2} \\ &= \frac{\sin(\pi\xi)}{\pi\xi};\end{aligned}$$

note how much cleaner the answer is in this case.

Additional problem (3). Recall a random variable has a chi-square distribution with d degrees of freedom if it has density

$$f_d(x) \ = \ \begin{cases}\frac{1}{2^{d/2}\Gamma(d/2)}x^{\frac{d}{2}-1}e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{otherwise,}\end{cases}$$

where Γ is the Gamma function (the generalization of the factorial function), which is given by

$$\Gamma(s) \ = \ \int_0^\infty x^{s-1}e^{-x}dx.$$

We know that if X_i has the standard normal distribution, then X_i^2 has the chi-square distribution with 1 degree of freedom. We write c_d for the normalization constant of the chi-square distribution with d degrees of freedom.

We first consider the case of the sum of two chi-square distributions, each with 1 degree of freedom. The density is

$$\begin{aligned}(f_1 * f_1)(x) &= \int_{-\infty}^\infty f_1(t)f_1(x-t)dt \\ &= \int_0^x c_1 t^{-1/2}e^{-t/2}\cdot c_1(x-t)^{-1/2}e^{-(x-t)/2}dt.\end{aligned}$$

The range of integration stops at x as $f_1(x-t)$ is zero if the argument is negative. Simplifying yields

$$(f_1 * f_1)(x) \;=\; c_1^2 e^{-x/2} \int_0^x t^{-1/2}(x-t)^{-1/2} dt.$$

There are two ways to proceed. The first is to try and evaluate this integral directly. It may be possible to do this through brute force, but it won't be pleasant. Note that the final answer *must* be a probability distribution. Thus, *we do not need to figure out the integral exactly; it suffices to determine the x dependence!* The reason is that if we know the x-dependence, then we get the normalization constant by integrating $(f_1 * f_1)(x)$ with respect to x and setting the result equal to 1.

Thus let us make the following clever change of variables: set $t = ux$ and $dt = xdu$; as t runs from 0 to x we have u runs from 0 to 1. This yields

$$\begin{aligned}(f_1 * f_1)(x) &= c_1^2 e^{-x/2} \int_0^1 (xu)^{-1/2}(x-xu)^{-1/2} x du \\ &= c_1^2 e^{-x/2} \frac{x}{x^{1/2}x^{1/2}} \int_0^1 u^{-1/2}(1-u)^{-1/2} du.\end{aligned}$$

The u-integral can be done in closed form, as it is proportional to integrating the Beta density (with parameters $\alpha = \beta = 1/2$); however, there is no need! Letting $\mathcal{C}_1$ denote the value of the u-integral, we see

$$(f_1 * f_1)(x) \;=\; \begin{cases} \mathcal{C}_1 c_1^2 e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{otherwise.}\end{cases}$$

For this to be a probability distribution, the integral must be 1, which implies $\mathcal{C}_1 c_1^2 = 1/2$. Again, we emphasize that while we could have computed $\mathcal{C}_1$ by brute force, there was no need. To show that we have a chi-square distribution with 2 degrees of freedom, it suffices to show that we have the correct x-dependence, as then the normalization constants must match.

We now turn to the general case. We proceed by induction. We have already handled the base case; now we must show $X_1^2 + \cdots + X_{n+1}^2$ is a chi-square distribution with $n+1$ degrees of freedom. By induction $X_1^2 + \cdots + X_n^2$ is a chi-square distribution with n degrees of freedom. Calling the normalization constants c_n and c_1 again, we see that

$$\begin{aligned}(f_1 * \cdots * f_1)(x) &= \int_{-\infty}^{\infty} f_n(t) f_1(x-t) dt \\ &= \int_0^x c_n t^{\frac{n}{2}-1} e^{-t/2} \cdot c_1 (x-t)^{-\frac{1}{2}} e^{-(x-t)/2} dt.\end{aligned}$$

The exponential factors combine to give $e^{-x/2}$, and we again set $t = ux$ and $dt = xdu$, and find

$$\begin{aligned}(f_1 * \cdots * f_1)(x) &= c_n c_1 e^{-x/2} \int_0^1 (xu)^{\frac{n}{2}-1}(x-xu)^{-\frac{1}{2}} x du \\ &= c_n c_1 e^{-x/2} x^{\frac{n}{2}-1} x^{-\frac{1}{2}} x \int_0^1 u^{\frac{n}{2}-1}(1-u)^{-\frac{1}{2}} du \\ &= c_n c_1 x^{\frac{n+1}{2}-1} e^{-x/2} \int_0^1 u^{\frac{n}{2}-1}(1-u)^{-\frac{1}{2}} du. \qquad (8.1)\end{aligned}$$

Again, it is possible to evaluate the u-integral in closed form (it is essentially the integral of a Beta density with parameters $n/2$ and $1/2$); however, all that matters is that it has no x-dependence. Calling this integral $\mathcal{C}_n$, we find

$$(f_1 * \cdots * f_1)(x) \ = \ \begin{cases} \mathcal{C}_n c_n c_1 x^{\frac{n+1}{2}-1} e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note the x-dependence is exactly that of the chi-square distribution with $n+1$ degrees of freedom, and thus the normalization constant $\mathcal{C}_n c_n c_1$ must equal the normalization constant of the chi-square with $n+1$ degrees of freedom. We emphasize again that we could have computed this constant by brute force, but that there was again no need!

For completeness, we state what the Beta density is. Let $u \in [0,1]$ and $\alpha, \beta > 0$. Then the Beta distribution with parameters α and β is given by the density

$$g_{\alpha,\beta}(u) \ = \ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}$$

for $u \in [0,1]$ and 0 otherwise. As this is a probability distribution, it integrates to 1 and thus

$$\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du \ = \ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

This is the integral we need if we want to do the integrals above.

9. HW #9

Due Thursday November 12 (though you may place in my mailbox anytime up till 10am on Friday 11/13): (1) Let $X_1, \dots, X_n$ be iidrv random variables with the geometric distribution with parameter p, so $\text{Prob}(X_i = k) = (1-p)^{k-1}p$ for k a positive integer and 0 otherwise. Let $\overline{X} = (X_1 + \cdots + X_n)/n$. Find $\mathbb{E}[\overline{X}]$, $\text{Var}(\overline{X})$, and the moment generating function of $Y = (\overline{X} - \mathbb{E}[\overline{X}])/\text{StDev}(\overline{X})$. (2) Calculate the Laplace transforms of the following densities (a) an exponential distribution with parameter λ; (b) uniform distribution on $[a,b]$ with $a \geq 0$. (3) For each function compute the complex derivative at $z = 0$ or prove the function is not differentiable there: (a) $f(z) = z$; (b) $f(z) = z^2$; (c) $f(z) = \overline{z}$, where if $z = x + iy$ then $\overline{z} = x - iy$. Recall that the derivative is defined by

$$f'(z) \ = \ \lim_{h \to 0} \frac{f(z+h) - f(z)}{h},$$

where $h = h_1 + ih_2$ tends to $0 + 0i$ along any path but is *never* 0 in any calculation. (4) Prove the product rule of differentiation, namely that if f and g are differentiable then the derivative of $f(x)g(x)$ is $f'(x)g(x) + f(x)g'(x)$. Using this, induction and the fact that the derivative of x is 1, compute the derivative of x^n for any positive integer n. Note that this proof *bypasses* having to use the binomial theorem to expand $(x+h)^n$! (5) Calculate the limits as $(x, y) \to (0, 0)$, or prove the limit does not exist: (a)

$$\lim_{(x,y)\to(0,0)} \frac{x^3 + 1701x^2y^2 + 24601y^4}{x^2+y^2};$$

(b)

$$\lim_{(x,y)\to(0,0)} \left[\frac{x^8+y^8}{x^2+y^8} - \frac{x^{10}+y^{10}}{x^4+y^{10}}\right].$$

(Extra Credit) Prove or disprove: notation as in the first problem, the MGF of Y converges to the MGF of the standard normal as n tends to infinity.

Problem 1. To compute the expected value, we use the expected value of a sum is the sum of the expected values. Thus

$$\begin{aligned}
\mathbb{E}[\overline{X}] &= \mathbb{E}[(X_1 + \cdots + X_n)/n] \\
&= \frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\
&= \frac{1}{n}\cdot n\frac{1}{p} \ = \ \frac{1}{p},
\end{aligned}$$

as the mean of a geometric random variable with parameter p is just $1/p$. While we have seen this result before, it is easily proved using moment generating functions; as we will need to work with these functions anyway for the rest of the problem, let's take a moment and rederive this result.

Let X be a random variable with the geometric distribution with parameter p. Then its moment generating function is

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{k=1}^{\infty} e^{tk}\text{Prob}(X=k) \\ &= \sum_{k=1}^{\infty} e^{tk}(1-p)^{k-1}p \\ &= \frac{p}{1-p}\sum_{k=1}^{\infty}(e^t(1-p))^k \\ &= \frac{p}{1-p}\sum_{k=0}^{\infty}(e^t(1-p))^k - \frac{p}{1-p} \\ &= \frac{p}{1-p}\frac{1}{1-(e^t(1-p))} - \frac{p}{1-p}. \end{aligned}$$

One of the most useful properties of moment generating functions is that $\mathbb{E}[X^\ell] = d^\ell M_X(t)/dt^\ell\Big|_{t=0}$; in particular, the mean is simply $\mathbb{E}[X] = M_X'(0)$, so

$$M_X'(t) \ = \ \frac{p}{1-p}\frac{e^t(1-p)}{(1-e^t(1-p))^2}, \quad M_X'(0) \ = \ \frac{1}{p}.$$

There are several ways of doing the algebra; we could use the formula for a geometric series starting at $n=1$ and not starting at $n=0$ which required us to subtract off the $n=0$ term. Why do we prefer this? The reason is that the resulting expression only has t dependence in the denominator; if we started the sum at $n=1$ we would have t dependence in both the numerator and denominator, which means we have to use the quotient rule to find $M_X'(t)$.

Knowing the moment generating function of X when $X \sim \text{Geom}(p)$ simplifies the remaining parts of the problem. For the variance, we have

$$\begin{aligned} \text{Var}(\overline{X}) &= \text{Var}((X_1+\cdots+X_n)/n) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) \\ &= \frac{1}{n^2}\cdot n\text{Var}(X) \\ &= \frac{1}{n^2}\cdot n\frac{1-p}{p^2} \ = \ \frac{1-p}{np^2}, \end{aligned}$$

where we used $\text{Var}(X) = (1-p)/p^2$. We can easily derive this from the moment generating function. As $\mathbb{E}[X^2] = M_X''(0)$, we have

$$\text{Var}(X) \ = \ \mathbb{E}[X^2] - \mathbb{E}[X]^2 \ = \ M_X''(0) - M_X'(0)^2.$$

We've already computed $M'_X(0)$, and thus only need to find $M''_X(0)$. We have

$$\begin{aligned}
M''_X(t) &= \frac{d}{dt}\left(\frac{p}{1-p}\frac{e^t(1-p)}{(1-e^t(1-p))^2}\right)\\
&= \frac{p}{1-p}\left(\frac{e^t(1-p)}{(1-e^t(1-p))^2}+\frac{2e^{2t}(1-p)^2}{(1-e^t(1-p))^3}\right)\\
M''_X(1) &= \frac{p}{1-p}\left(\frac{1-p}{p^2}+\frac{2(1-p)^2}{p^3}\right)\\
&= \frac{1}{p}+\frac{2(1-p)}{p^2}\\
&= \frac{p+2-2p}{p^2} \ = \ \frac{2-p}{p^2}.
\end{aligned}$$

Therefore the variance is

$$\mathrm{Var}(X) \ = \ M''_X(0)-M'_X(0)^2 \ = \ \frac{2-p}{p^2}-\frac{1}{p^2} \ = \ \frac{1-p}{p^2}.$$

The last part of the problem asks us to compute the moment generating function of $Y=(\overline{X}-\mathbb{E}[\overline{X}])/\mathrm{StDev}(\overline{X})$. Let

$$\begin{aligned}
b &= \mathbb{E}[\overline{X}] \ = \ \frac{1}{p}\\
a &= \mathrm{StDev}(\overline{X}) \ = \ \sqrt{\frac{1-p}{np^2}} \ = \ \frac{\sqrt{1-p}}{p\sqrt{n}}.
\end{aligned}$$

In the arguments below, we constantly use $M_{\alpha W+\beta}(t)=e^{\beta t}M_W(\alpha t)$. We have

$$\begin{aligned}
M_Y(t) &= M_{(\overline{X}-b)/a}(t)\\
&= M_{\overline{X}/a-b/a}(t)\\
&= e^{-bt/a}M_{\overline{X}}(t/a)\\
&= e^{-bt/a}M_{(X_1+\cdots+X_n)/n}(t/a)\\
&= e^{-bt/a}M_{X_1+\cdots+X_n}\left(\frac{t}{an}\right)\\
&= e^{-bt/a}M_{X_1}\left(\frac{t}{an}\right)\cdots M_{X_n}\left(\frac{t}{an}\right)\\
&= e^{-bt/a}M_X\left(\frac{t}{an}\right)^n.
\end{aligned}$$

Earlier, however, we showed that

$$M_X(t) \ = \ \frac{p}{1-p}\frac{1}{1-(e^t(1-p))}-\frac{p}{1-p}.$$

Substituting yields

$$M_Y(t) \ = \ e^{-bt/a}\left(\frac{p}{1-p}\right)^n\left(\frac{1}{1-(e^{t/an}(1-p))}-1\right)^n,$$

with

$$\begin{aligned} b &= \mathbb{E}[\overline{X}] = \frac{1}{p} \\ a &= \text{StDev}(\overline{X}) = \sqrt{\frac{1-p}{np^2}} = \frac{\sqrt{1-p}}{p\sqrt{n}}. \end{aligned}$$

Problem 2. We first compute the Laplace transform of the standard exponential, which has density function $f(x) = e^{-x}$ for $x \geq 0$ and 0 otherwise. We have

$$\begin{aligned} (\mathcal{L}f)(s) &= \int_0^\infty f(t)e^{-st}dt \\ &= \int_0^\infty e^{-t}e^{-st}dt \\ &= \int_0^\infty e^{-(1+s)t}dt \\ &= \frac{1}{1+s}\int_0^\infty e^{-(1+s)t}(1+s)dt \\ &= \frac{1}{1+s}, \end{aligned}$$

so long as $s > -1$ (we need this in order to ensure that the argument of the exponential is negative, as otherwise the integral diverges).

We now compute the Laplace transform of the uniform distribution on $[a, b]$ with $a \geq 0$. The density function is $f(x) = \frac{1}{b-a}$ if $a \leq x \leq b$ and 0 otherwise. Thus

$$\begin{aligned} (\mathcal{L}f)(s) &= \int_0^\infty f(t)e^{-st}dt \\ &= \int_a^b \frac{1}{b-a}e^{-ts}dt \\ &= \frac{1}{b-a}\int_a^b e^{-ts}dt \\ &= \frac{1}{(b-a)s}\int_a^b e^{-ts}sdt \\ &= \frac{-1}{(b-a)s}\left(e^{-bs} - e^{-as}\right) \\ &= \frac{e^{-as} - e^{-bs}}{(b-a)s}. \end{aligned}$$

Problem 3. While the problem only asks whether or not the functions are differentiable at $z = 0$ (and if so what the derivative is), we consider the more general case as the argument is essentially the same. We let $h = h_1 + ih_2$ below, with $h \to 0 + 0i$. For (a), we have $f(z) = z$ so

$$\lim_{h\to 0} \frac{f(z+h)-f(z)}{h} \ = \ \lim_{h\to 0} \frac{z+h-z}{h} \ = \ \lim_{h\to 0} 1 \ = \ 1;$$

thus the function is complex differentiable and the derivative is 1.

For (b), we have $f(z) = z^2$ and

$$\begin{aligned}\lim_{h\to 0} \frac{f(z+h)-f(z)}{h} \ &= \ \lim_{h\to 0} \frac{(z+h)^2-z^2}{h} \\ &= \ \lim_{h\to 0} \frac{z^2+2zh+h^2-z^2}{h} \\ &= \ \lim_{h\to 0} \frac{2zh+h^2}{h} \\ &= \ \lim_{h\to 0}(2z+h) \\ &= \ \lim_{h\to 0} 2z + \lim_{h\to 0} h \\ &= \ 2z+0 \ = \ 2z.\end{aligned}$$

We are using the following properties of complex numbers: $h/h = 1$ and $2zh + h^2 = (2z+h)h$.

For (c), we have $f(z) = \overline{z}$, and thus

$$\lim_{h\to 0} \frac{f(z+h)-f(z)}{h} \ = \ \lim_{h\to 0} \frac{\overline{z+h}-\overline{z}}{h}.$$

Unlike the other limits, this one is not immediately clear. Let us write $z = x + iy$, $h = h_1 + ih_2$ (and of course $\overline{z} = x - iy$, $\overline{h} = h_1 - ih_2$). We therefore find the limit is

$$\lim_{h\to 0} \frac{x-iy+h-ih_2-(x-iy)}{h_1+ih_2} \ = \ \lim_{h\to 0} \frac{h_1-ih_2}{h_1+ih_2}.$$

This limit does not exist; depending on how $h \to 0$ we obtain different answer. For example, if $h_2 = 0$ (traveling along the x-axis) the limit is just $\lim_{h\to 0} h_1/h_1 = 1$, while if $h_1 = 0$ (traveling along the y-axis) the limit is just $\lim_{h\to 0} -ih_2/ih_2 = -1$. Thus this function is not complex differentiable anywhere.

If we continue to argue along these lines, we find that a function is complex differentiable if the x and y dependence is in a very special form, namely everything is a function of $z = x + iy$. In other words, we do not allow our function to depend on $\overline{z} = x - iy$. If we could depend on both, we could isolate out x (which is $z + \overline{z}$) and y (which is $(z - \overline{z})/i$). We can begin to see why being complex differentiable once implies that we are complex differentiable infinitely often, namely because of the very special dependence on x and y.

Problem 4. Let $A(x) = f(x)g(x)$. Then

$$\begin{aligned}
A'(x) &= \lim_{h\to 0} \frac{A(x+h) - A(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h)\textbf{-f(x)g(x+h)+f(x)g(x+h)} - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \left[\frac{f(x+h) - f(x)}{h}\, g(x+h) + f(x)\frac{g(x+h) - g(x)}{h}\right] \\
&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \lim_{h\to 0} g(x+h) + \lim_{h\to 0} f(x) \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} \\
&= f'(x)g(x) + f(x)g'(x).
\end{aligned}$$

We proceed by induction to prove the derivative of x^n is nx^{n-1} for n a positive integer. The base case is clear (and we are in fact told we may assume this). Thus we are left with proving the inductive step, namely given that the derivative of x^n is nx^{n-1} we must prove the derivative of x^{n+1} is $(n+1)x^n$. Let $f(x) = x^n$ and $g(x) = x$. We use the product rule; by induction $f'(x) = nx^{n-1}$ (and of course $g'(x) = 1$). The product rule tells us that

$$(x^{n+1})' \ = \ f'(x)g(x) + f(x)g'(x) \ = \ nx^{n-1} \cdot x + x^n \cdot 1 \ = \ (n+1)x^n.$$

Thus, although initially it appears that we need the binomial theorem to compute the derivative of x^n, we can actually bypass it by using induction and the product rule!

As an aside, the derivative of x^r for general $r \in \mathbb{R}$ cannot be obtained by arguments such as those above. We can find the derivative of $h(x) = x^{p/q}$ by using the power rule applied to $A(x) = h(x)^q = x^p$, and then solving for $h'(x)$. The algebra starts with

$$A'(x) \ = \ qh(x)^{q-1}h'(x) \ = \ px^{p-1}, \quad h(x) \ = \ x^{p/q}.$$

We now isolate $h'(x)$, and find

$$h'(x) \ = \ \frac{px^{p-1}}{qh(x)^{q-1}} \ = \ \frac{p}{q}\,\frac{x^{p-1}}{x^{(p/q)(q-1)}} \ = \ \frac{p}{q}\,x^{p-1-\frac{p(q-1)}{q}} \ = \ \frac{p}{q}\,x^{\frac{p}{q}-1}.$$

For general r, we need to write $x^r = \exp(r \ln x)$ and use the chain rule. Why do we need to do this? We can interpret $(x+h)^n$ when n is an integer, but what does it mean to have $(x+h)^{\sqrt{2}}$?

Problem 5. For (a), the limit is zero. The easiest way to see this is to convert to polar coordinates, with $x = r\cos\theta$ and $y = r\sin\theta$. As $(x,y) \to (0,0)$, we have $r \to 0$. The numerator is bounded by $r^3 + 1701r^4 + 24601r^4$, while the denominator is r^2. Thus the ratio of the numerator over the denominator is at most $r + 26302r^2$, which tends to zero as $r \to 0$. For (b), we must consider the following limit:

$$\lim_{(x,y)\to(0,0)} \left[\frac{x^8 + y^8}{x^2 + y^8} - \frac{x^{10} + y^{10}}{x^4 + y^{10}}\right].$$

Let's check some special paths $(f(h), g(h))$ with $h \to 0$ to build intuition. We always assume $a \neq 0$ below.

x	y	Difference of Ratios	Limit as $h \to 0$
h	0	$h^6 - h^6$	0
0	h	$1 - 1$	0
h	ah	$\frac{(1+a^8)}{1+a^8h^6}h^6 - \frac{(1+a^{10})}{1+a^{10}h^6}h^6$	0
h^2	0	$h^{12} - h^{12}$	0
0	h^2	$1 - 1$	0
h	ah^2	$\frac{1+a^8h^8}{1+a^8h^{12}}h^6 - \frac{1+a^{10}h^{10}}{1+a^{10}h^{12}}h^6$	0
ah^2	h	$\frac{a^{16}h^8+1}{a^4+h^4}h^4 - \frac{a^{20}h^{10}+1}{a^8+h^8}h^{12}$	0

The evidence sure seems to suggest that the limit is zero. It is zero if we approach the origin along any line containing the origin, or on any pure parabola $y = mx^2$ or $x = my^2$, as well as quadratically decaying along the x or y-axes. Unfortunately, we cannot prove a limit exists by checking a fixed number of paths; we can only prove the limit exists by checking *all possible paths*, or by finding a path where the limit does not exist. For example, we must also consider the path in Figure 1.

It turns out that, if we investigate cubic paths, we see the limit does not exist. Specifically, consider the path $x = y^3$, or more specifically, $(x, y) = (h^3, h)$. This leads to

$$\lim_{h \to 0} \left[\frac{h^{24} + h^8}{h^6 + h^8} - \frac{h^{30} + h^{10}}{h^{12} + h^{10}} \right];$$

the first term looks like $h^8/h^6 = h^2$ for small h, while the second looks like $h^{10}/h^{10} = 1$. Thus the limit along this path is 1, which does not equal the previous limits of zero; thus this function does not have a limit as $(x, y) \to (0, 0)$.

We leave it as a fun exercise to the reader to think about how this strange example was generated, and to come up with a related example that has a limit among cubics but not among quartics.

Remember that for $|h| < 1$, $|h^n| > |h^m|$ if $n < m$ (for example, $|h^4| > |h^8|$). Thus for small h the numerators and denominators are controlled by the smaller powers of h. One way we can analyze these quantities is factoring:

$$\lim_{h \to 0} \frac{h^{24} + h^8}{h^6 + h^8} = \lim_{h \to 0} \frac{h^8}{h^6} \frac{h^{16} + 1}{1 + h^2} = \lim_{h \to 0} h^2 \frac{1 + h^{16}}{1 + h^2} = \lim_{h \to 0} h^2 \lim_{h \to 0} \frac{1 + h^{16}}{1 + h^2} = 0 \cdot 1,$$

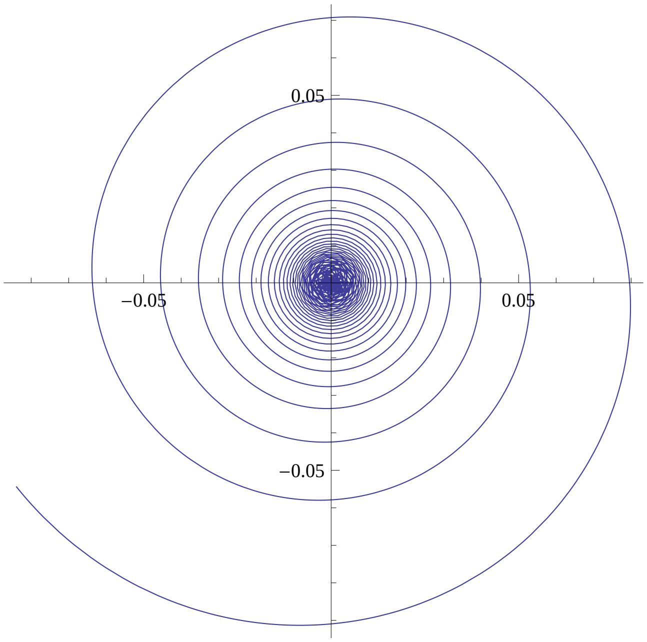

FIGURE 1. Path $(r\cos(1/r), r\sin(1/r))$ as $r \to 0$. If the limit exists, then we must get the same limit along this path as along the more standard paths such as straight lines, quadratics, et cetera.

where we use the limit of a product is the product of the limits (provided at least one exists).

The behavior in this problem is strange – the limit exists and is zero along any straight line or the standard parabolas, but not along a cubic. How can we reconcile this? The explanation is as follows: while the limit is zero along each straight line, the rate of convergence depends on the steepness of the line. In particular, if we go along the path $x = y^3$, we cut through these lines so quickly that we see a different behavior. A plot helps; see Figure 2.

To try and get a sense, we look at how the limit exists along various lines in Figure 3. Looking at these plots, we can see the difference in behavior, and if we choose a certain path $(x, g(x))$ we won't have a limit of zero.

Extra Credit Problem. The moment generating function of the standard normal is e^{t^2}, so the logarithm of the standard normal's moment generating function is t^2. Knowing this, it is natural to try and show that $\log M_Y(t) \to t^2$ as $n \to \infty$. Another reason why it is natural to look at logarithms is that $M_Y(t)$ involves factors to the n^{th} power, and

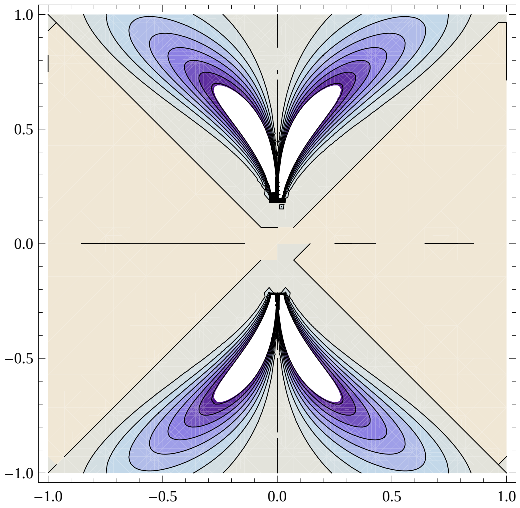

FIGURE 2. Contour plot of $\frac{x^8+y^8}{x^2+y^8} - \frac{x^{10}+y^{10}}{x^4+y^{10}}$.

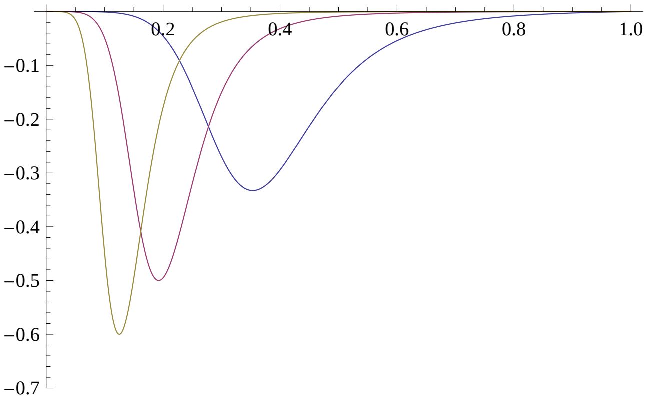

FIGURE 3. Plots of $g(x,y) = \frac{x^8+y^8}{x^2+y^8} - \frac{x^{10}+y^{10}}{x^4+y^{10}}$ along the lines $(h,2h)$, $(h,3h)$ and $(h,4h)$. The x-axis is h and the y-axis is the value of g. We can see that while these paths all have the same limit, they approach that limit differently.

taking logarithms brings down the n. We have

$$\begin{aligned}
\log M_Y(t) &= \log\left[e^{-bt/a}\left(\frac{p}{1-p}\right)^n\left(\frac{1}{1-(e^{t/an}(1-p))}-1\right)^n\right]\\
&= -\frac{bt}{a}+n\log\frac{p}{1-p}+n\log\left[\frac{1}{1-(e^{t/an}(1-p))}-1\right]\\
&= -\frac{bt}{a}+n\log\frac{p}{1-p}+n\log\left[\frac{e^{t/an}(1-p)}{1-(e^{t/an}(1-p))}\right]\\
&= -\frac{bt}{a}+n\log\frac{p}{1-p}+n\frac{t}{an}+n\log(1-p)-n\log\left[1-(e^{t/an}(1-p))\right]\\
&= -\frac{bt}{a}+n\log p+\frac{t}{a}-n\log\left[1-(e^{t/an}(1-p))\right]\\
&= -\frac{t}{ap}+\frac{t}{a}+n\log p-n\log\left[1-(e^{t/an}(1-p))\right]\\
&= -\frac{1-p}{p}\frac{t}{a}+n\log p-n\log\left[1-(e^{t/an}(1-p))\right].
\end{aligned}$$

Note $an = \frac{\sqrt{1-p}}{p}\sqrt{n}$, so for n large $e^{t/an}$ is approximately 1; in fact, we have

$$e^{t/an} \ = \ 1 + \frac{t}{an} + \frac{t^2}{2(an)^2} + O\left(\frac{t^3}{n^3}\right),$$

where $O(z)$ means there is a universal constant C such that the error is at most Cz. The reason we stop the expansion here is that we multiply the logarithm by n; once we have an error of size $O(1/n^{3/2})$ or smaller it will be dwarfed in the limit.

Using $\log(1-u) = -u - u^2/2 + O(u^3)$ we have

$$\begin{aligned} n\log\left[1-(e^{t/an}(1-p))\right] &= n\log\left[1-(1-p)\left(1+\frac{t}{an}+\frac{t^2}{2(an)^2}+O\left(\frac{t^3}{n^{3/2}}\right)\right)\right] \\ &= n\log\left[p-(1-p)\left(\frac{t}{an}+\frac{t^2}{2(an)^2}+O\left(\frac{t^3}{n^{3/2}}\right)\right)\right] \\ &= n\log p + n\log\left[1-\frac{1-p}{p}\left(\frac{t}{an}+\frac{t^2}{2(an)^2}+O\left(\frac{t^3}{n^{3/2}}\right)\right)\right] \\ &= n\log p - n\frac{1-p}{p}\left(\frac{t}{an}+\frac{t^2}{2(an)^2}\right) - \frac{n}{2}\left(\frac{1-p}{p}\frac{t}{an}\right)^2 + O\left(\frac{t^3}{n^{1/2}}\right) \\ &= n\log p - \frac{1-p}{p}\frac{t}{a} + \frac{1-p}{p}\left(1+\frac{1-p}{p}\right)\frac{t^2}{2a^2n} + O\left(\frac{t^3}{n^{1/2}}\right) \\ &= n\log p - \frac{1-p}{p}\frac{t}{a} + \frac{1-p}{p^2}\frac{t^2}{2a^2n} + O\left(\frac{t^3}{n^{1/2}}\right) \\ &= n\log p - \frac{1-p}{p}\frac{t}{a} + \frac{t^2}{2} + O\left(\frac{t^3}{n^{1/2}}\right), \end{aligned}$$

where we used the definition of a in the final step. Substituting this into $\log M_Y(y)$ we see most of the terms cancel, yielding

$$\log M_Y(t) \ = \ \frac{t^2}{2} + \left(\frac{t^3}{n^{1/2}}\right).$$

As $n \to \infty$ for any fixed t this converges to $t^2/2$. Thus as $n \to \infty$ we have $\log M_Y(t) \to t^2/2$, implying that $M_Y(t) \to e^{t^2/2}$ as claimed.

It is worth noting that while we were able to prove the claim, the above algebra is quite long and tedious and not at all enlightening. While this is essentially a proof of the Central Limit Theorem in this special case, the final result seems almost miraculous.

10. HW #10

Due Thursday November 19 (though you may place in my mailbox up till 10am on Friday 11/20):

(1) Let $X_1, \dots, X_N$ be iidrv with mean 0 and variance 1, and let $\overline{X} = (X_1 + \cdots + X_N)/N$. For any fixed $\epsilon > 0$, prove that as $N \to \infty$ we have

$$\lim_{N\to\infty} \text{Prob}(|\overline{X} - 0| > \epsilon) \ = \ 0.$$

Hint: try using Chebyshev's theorem.

(2) Let

$$f(x) \ = \ \begin{cases} \frac{1}{\sqrt{2\pi x^2}}\, e^{-(\log x)^2/2} & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

(a) Prove f is continuous.
(b) Let m be any fixed positive integer. Prove that $\lim_{x\to\infty} x^m f(x) = 0$.

(3) Assume you have a table of probabilities of the standard normal random variable X; in other words, you can easily look up probabilities of the following form: $\Phi(x) = \text{Prob}(X \le a)$. (The cumulative distribution function of the standard normal is used so often it gets a symbol reserved for it, namely Φ.)
(a) Show $\text{Prob}(X \le 0) = 1/2$.
(b) Let $Y \sim N(\mu, \sigma)$ be a normal random variable with mean μ and variance σ^2. Express $\text{Prob}(Y \le a)$ in terms of Φ, μ, σ and of course a.

(4) **DO EXACTLY ONE OF THE FOLLOWING:**

(a) Find any math research paper or expository paper which uses probability and write an at most one page summary (preferably in TeX). As you continue in your careers, you are going to need to read technical papers and summarize them to your superiors / colleagues / clients; this is thus potentially a very useful exercise. Make sure you describe clearly what the point of the paper is, what techniques are used to study the problem, what applications there are (if any). Below is a sample review from MathSciNet; if you would like to see more, you can go to their homepage or ask me and I'll pass along many of the ones I've written. I've chosen this one as it's related to a paper on randomly shuffling cards (this paper is linked in the additional comments from October 27): Bayer, Dave and Diaconis, Persi, *Trailing the dovetail shuffle to its lair,* Ann. Appl. Probab. **2** (1992), no. 2, 294–313.

Rarely does a new mathematical result make both the New York Times and the front page of my local paper, and even more rarely is your reviewer asked to speak on commercial radio about a result, but such activity was caused by the preprint of this paper. In layman's terms, it says you should shuffle a deck of cards seven times before playing. More technically, the usual way people shuffle is called a riffle shuffle, and a natural mathematical model of a random *shuffle is to assume all possible riffle shuffles are equally likely. With this model one can ask how close is k shuffles of an n-card deck to the uniform distribution on all $n!$ permutations, where 'close' is measured by variation distance. It was previously known that, as $n \to \infty$, one needs $k(n) \sim 32\log_2 n$ shuffles to get close to uniform. This paper gives an elegant and careful treatment based on an explicit formula for the exact distance $d(k,n)$ to uniformity. To quote the abstract: 'Key ingredients are the analysis of a card trick and the determination of the idempotents of a natural commutative subalgebra in the symmetric group algebra.' – Reviewed by David J. Aldous*

(b) The following three problems: Problem #10 in Section 3.11 **and** Problem #17 in Section 3.11 **and** Problem #1 in Section 1.8.

(5) **Extra Credit:** Prove which of the following from lecture converges slowest to the standard normal: uniform, Laplace or Millered Cauchy.

(6) **Extra Credit:** Define

$$f_k(x) \ = \ \frac{C_k(a)}{1+(ax)^{2k}}$$

where $C_k(a)$ is chosen so that the above is a probability density.
(a) Find a and $C_3(a)$ so that the density above has variance 1.
(b) More generally, for any integer $k \geq 3$ find a and $C_k(a)$ so that the density above has variance 1.

Problem 1. Chebyshev's theorem states that if a random variable Y has finite mean μ_Y and finite variance σ_Y^2, then $\text{Prob}(|Y-\mu_Y| > k\sigma) \leq 1/k^2$. Let's take $Y = \overline{X} = (X_1 + \cdots + X_N)/N$. Then $\mathbb{E}[Y] = 0$ (as each X_i has mean 0) and $\text{Var}(Y) = \frac{1}{N^2}\sum_{i=1}^N \text{Var}(X_i) = 1/N$, so $\sigma_Y = 1/\sqrt{N}$.

We must determine how many standard deviations ϵ is. As one standard deviation is $1/\sqrt{N}$, it takes $k = \epsilon\sqrt{N}$ standard deviations to be at least ϵ away from the mean. As ϵ is fixed and $N \to \infty$, eventually this number is arbitrarily large. Thus by Chebyshev we find

$$\text{Prob}(|\overline{X}-0| > \epsilon) \ = \ \text{Prob}\left(|\overline{X}-0| > \epsilon\sqrt{N}\sigma_Y\right) \ \leq \ \frac{1}{\epsilon^2 N},$$

which clearly tends to zero for any fixed ϵ as $N \to \infty$.

*It is worth remarking that Problem 1 is actually a very famous and very important theorem, namely the **Weak Law of Large Numbers**. The Central Limit Theorem is a strengthening of this, where in addition to knowing $\overline{X}_N \to \mu$ we know how it converges as well.*

Problem 2. We first note that this is not a randomly chosen function. This is one of the two densities we saw earlier that have the same integral moments but are different.

(a) To prove that f is continuous, it clearly suffices to check the behavior of f as $x \to 0$. We thus need to show $\lim_{x\to 0} f(x) = 0$. To evaluate this limit, it is natural to try and use L'Hopital's rule. I prefer to move the $\exp(-(\log x)^2/2)$ to the denominator and let $y = 1/x$; this gives a nice ratio of ∞/∞ with the variable tending to infinity. Thus we must evaluate

$$\lim_{y\to\infty} \frac{y}{\sqrt{2\pi}\exp((\log y)^2/2)}.$$

If we try and use L'Hopital's rule, we run into some difficulties as the derivative of the exponential factor is $\exp((\log y)) \cdot \log y \cdot \frac{1}{y}$, which gives

$$\lim_{y\to\infty} \frac{y}{\sqrt{2\pi}\exp((\log y)^2/2)} \ = \ \lim_{y\to\infty} \frac{y}{\sqrt{2\pi}\exp((\log y)^2/2)\cdot\log y}.$$

We can surmount this by changing variables yet again, setting $y = e^w$ or $w = \log y$. Thus we must study

$$\lim_{w\to\infty} \frac{e^w}{\sqrt{2\pi}e^{w^2/2}}.$$

There is actually no need to apply L'Hopital's rule, as we can compute the limit directly. It is

$$\lim_{w\to\infty} \frac{1}{\sqrt{2\pi}\exp(w(\frac{w}{2}-1))} \ = \ 0.$$

For another approach, we again change variables with $y = 1/x \to \infty$ and find our limit equals

$$\begin{aligned}\lim_{y\to\infty} \frac{y}{\sqrt{2\pi}}\frac{1}{e^{(\log y)^2/2}} \ &= \ \frac{1}{\sqrt{2\pi}}\lim_{y\to\infty}\frac{y}{e^{(\log y)^2/2}} \\ &= \ \frac{1}{\sqrt{2\pi}}\lim_{y\to\infty}\frac{y}{\left(e^{\log y}\right)^{(\log y)/2}} \\ &= \ \frac{1}{\sqrt{2\pi}}\lim_{y\to\infty}\frac{y}{y^{(\log y)/2}} \ = \ 0.\end{aligned}$$

As $\log y \to \infty$, for y large the denominator is clearly greater than y^{341}, which suffices to prove our claim. It is very natural to approach the problem this way. The reason is that we are happy when an exponential hits a logarithm, as the two functions are inverses and cancel. Thus we want to exploit the fact that we are exponentiating a polynomial in the logarithm of x.

Another way to attack this problem is to study the limit as $x \to 0$ of the logarithm of our quantity; if this limit tends to negative infinity then the limit of the original quantity tends to the exponential of negative infinity, or zero. What is the limit of the logarithm? It is just

$$\begin{aligned}\lim_{x\to 0}\log\left[\frac{1}{\sqrt{2\pi x^2}}\,e^{-(\log x)^2/2}\right] \ &= \ \lim_{x\to 0}\left[\log x + \frac{\log 2\pi}{2} - \frac{(\log x)^2}{2}\right] \\ &= \ \frac{\log 2\pi}{2} - \lim_{x\to 0}\log x\cdot\left(\frac{\log x}{2} - 1\right).\end{aligned}$$

As $x \to 0$, both factors involving logarithms tend to minus infinity, and thus their product tends to infinity; as we multiply by negative one, the limit above is minus infinity, and hence the original limit is just zero.

For (b), we want to prove that $\lim_{x\to\infty} x^m f(x) = 0$. Let's analyze the factor $e^{(\log x)^2/2}$. Arguing as in the alternative proof for the previous part, we find

$$e^{(\log x)^2/2} \ = \ \left(e^{\log x}\right)^{\frac{\log x}{2}} \ = \ (x)^{\frac{\log x}{2}} \ = \ \sqrt{x}^{\log x}.$$

Thus

$$x^m f(x) \ = \ \frac{x^m}{\sqrt{2\pi x^2}\sqrt{x}^{\log x}}.$$

If we take x so large that $\log x > 4m$, then $\sqrt{x}^{\log x} > x^{2m}$, and thus the denominator grows faster than the numerator, so the limit is zero. Specifically,

$$x^m f(x) \ = \ \frac{x^m}{\sqrt{2\pi x^2}\sqrt{x}^{\log x}} \ > \ \frac{x^m}{\sqrt{2\pi x^2}x^{2m}} \ = \ \frac{1}{\sqrt{2\pi x^2}x^m},$$

which clearly tends to zero as $x \to \infty$.

Problem 3. (a) The density of the standard normal is symmetric about $x = 0$ (in other words, it is an even function), as it is just

$$f(x) \ = \ \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}.$$

As the density is symmetric about 0, we have

$$\Phi(0) \ = \ \int_{-\infty}^{0} f(x)dx \ = \ \int_{0}^{\infty} f(x)dx;$$

as the sum of these two integrals must equal 1, each one is therefore equal to 1/2.

The reason this problem is so important is that there are three different tables of probabilities for the standard normal, and noting the symmetry is useful in converting from one to another. Specifically, we could be given the probability from $-\infty$ to x (which is $\Phi(x)$), the probability from $-x$ to x (which is $\Phi(x) - \Phi(-x)$), or the probability from 0 to x (which is $\Phi(x) - \Phi(0) = \Phi(x)$).

(b) As $Y \sim N(\mu, \sigma)$, we have $Z = (Y - \mu)/\sigma \sim N(0, 1)$. To see this, note that clearly Z is normally distributed, and we've adjusted everything so that Z has mean 0 and variance 1, implying Z is the standard normal. We can solve for Y in terms of Z, and find $Y = \sigma Z + \mu$. We therefore have

$$\begin{aligned}
\text{Prob}(Y \le a) &= \text{Prob}\left(\sigma Z + \mu \le a\right) \\
&= \text{Prob}(\sigma Z \le a - \mu) \\
&= \text{Prob}\left(Z \le \frac{a-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{a-\mu}{\sigma}\right)
\end{aligned}$$

11. HW #11

Due Thursday November 19. **DO ANY FIVE OF THE PROBLEMS BELOW.** *If you choose to do either problem 6 or 7 (you of course may elect to do both), you must email me your .tex file and .pdf, and note on the homework you submit to the grader which of these problems you elected to do.*

(1) Let $X, X_1, \dots, X_N$ be independent exponential random variables with parameter λ. Find the moment generating function for X_i. Directly using the moment generating function, prove the central limit theorem for $X_1 + \cdots + X_N$ (i.e., mimic what we did for the Poisson).

(2) Let $f(x)$ be a Schwartz function on $(-\infty, \infty)$. In particular, this means that f is a k times continuously differentiable probability density for any positive integer j. In other words, the first k derivatives of f exist and each of these derivatives is continuous. Prove there is some constant C (depending on f, f', $\dots$, $f^{(k)}$ such that as $|y| \to \infty$, $|\widehat{f}(y)| \le C/|y|^k$.

(3) We say f is a continuous probability density supported on $[-B, B]$ if $f(x) = 0$ if $|x| > B$; equivalently if X is a random variable with density f we say X is supported on $[-B, B]$ if f is supported on $[-B, B]$. For example, if $X \sim \mathrm{Unif}(2,5)$ then X is supported on $[-5, 5]$, while if $X \sim \mathrm{Exp}(1)$ then there is no B such that X is supported on $[-B, B]$.

- ◇ Prove or disprove: if f is supported on $[-B, B]$ then the $2k^{\text{th}}$ moment of f is at most B^{2k}.
- ◇ Prove or disprove: Let μ'_{2k} denote the $2k^{\text{th}}$ moment of f. Assume that
$$\lim_{k\to\infty} \mu'^{1/2k}_{2k} \ \le\ B.$$
Then f is supported on $[-B, B]$. (In other words, the probability of f taking on a value x with $|x| > B$ is zero.)
- ◇ Prove or disprove: Assume μ'_{2k}, the $2k^{\text{th}}$ moment of f, satisfies
$$(2k)!! \ \le\ \mu'_{2k} \ \le\ (2k)!.$$
Then there is some finite B such that f is supported on $[-B, B]$.

(4) Is the following argument correct: Consider
$$\lim_{N\to\infty} \left[\left(1 + \frac{x}{N}\right)^{N^2} \cdot \left(1 - \frac{x}{N}\right)^{N^2}\right].$$
For large N the first factor looks like e^{xN} since
$$\left(1 + \frac{x}{N}\right)^{N^2} \ =\ \left(\left(1 + \frac{x}{N}\right)^{N}\right)^{N} \ \longrightarrow\ (e^x)^N \ =\ e^{xN}.$$

Similarly we see that the second factor looks like e^{-xN}, and thus the product tends to 1 as $N \to \infty$. If this argument is wrong, what should the limit be? In other words, find

$$\lim_{N\to\infty}\left[\left(1+\frac{x}{N}\right)^{N^2}\cdot\left(1-\frac{x}{N}\right)^{N^2}\right]$$

if the argument above is incorrect.

(5) Let A be an arithmetic progression of n integers with common difference d; this means there is some n_0 such that

$$A \ = \ \{n_0, n_0+d, n_0+2d, \dots, n_0+nd\}.$$

Prove $|A+A| = |A-A|$, where

$$\begin{aligned} |A+A| &= \{a_1+a_2 : a_1, a_2 \in A\} \\ |A-A| &= \{a_1-a_2 : a_1, a_2 \in A\}. \end{aligned}$$

This implies that arithmetic progressions are 'balanced' (their sumset $A+A$ is as large as their difference set $A-A$). *Hint: show without loss of generality that we may take $n_0 = 0$ and $d = 1$ when we count the number of differences or sums.*

(6) Write up a problem of your choosing and a solution. You must have someone from the class check it. If the problem is unclear or the solution is wrong, unlike previous homework assignments this time you will lose points.

(7) Read a paper involving probability and give a one page summary.

Problem 1. We first compute the moment generating function of exponential random variables with parameter λ. It is

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= \int_0^\infty e^{tx}e^{-\lambda x}\lambda dx \\ &= \frac{\lambda}{\lambda-t}\int_0^\infty e^{-(\lambda-t)x}(\lambda-t)dx \\ &= \left(1-\frac{t}{\lambda}\right)^{-1}. \end{aligned}$$

There are many ways to do algebra; we chose to multiply by 1 in the form of $\frac{\lambda-t}{\lambda-t}$ as the exponential's argument is $-(\lambda-t)x$. In other words, we essentially have an exponential with parameter $\lambda-t$, and thus we just need to do some algebra to get the right density, which integrates to 1. Multiplying by $\frac{\lambda-t}{\lambda-t}$ leads to integrating $\exp(-(\lambda-t)x)(\lambda-t)$, which is an exponential with parameter $\lambda-t$.

The Central Limit Theorem involves studying the limiting distribution of

$$Z_N \;=\; \frac{\overline{X}-\mu}{\sigma/\sqrt{N}} \;=\; \sum_{n=1}^{N} \frac{X_i-\mu}{\sigma\sqrt{N}}.$$

As the moment generating function of a sum of independent random variables is the product of the moment generating functions, we have

$$M_{Z_N}(t) \;=\; \prod_{n=1}^{N} e^{\frac{-\mu t}{\sigma\sqrt{N}}}\, M_X\left(\frac{t}{\sigma\sqrt{N}}\right) \;=\; e^{\frac{-\mu t\sqrt{N}}{\sigma}} M_X\left(\frac{t}{\sigma\sqrt{N}}\right)^N.$$

Taking logarithms we find

$$\log M_{Z_N}(t) \;=\; -\frac{\mu t\sqrt{N}}{\sigma} + N\log M_X\left(\frac{t}{\sigma\sqrt{N}}\right).$$

We now stop arguing in full generality and instead use the fact that we have a sum of exponential random variables with parameter λ. The mean and the standard deviation are both $1/\lambda$, so

$$\log M_{Z_N}(t) \;=\; -t\sqrt{N} + N\log M_X\left(\frac{\lambda t}{\sqrt{N}}\right).$$

Substituting for the moment generating function yields

$$\log M_{Z_N}(t) \;=\; -t\sqrt{N} - N\log\left(1-\frac{t}{\sqrt{N}}\right).$$

We Taylor expand, using

$$\log(1-u) \;=\; -\left(u+\frac{u^2}{2}+\cdots\right)$$

and find

$$\begin{aligned}\log M_{Z_N}(t) \quad &= \quad -t\sqrt{N} + N\left(\frac{t}{\sqrt{N}}+\frac{1}{2}\frac{t^2}{N}+O\left(\frac{1}{N^{3/2}}\right)\right)\\ &= \quad \frac{t^2}{2}+O\left(\frac{1}{N^{1/2}}\right),\end{aligned}$$

which implies

$$M_{Z_N}(t) \;=\; e^{t^2/2}e^{O(N^{-1/2})},$$

which converges to the moment generating function of the standard normal as $N\to\infty$. Using our results from complex analysis, the fact that the moment generating functions exist in a neighborhood of the origin and that they converge to the moment generating function of the standard normal, we now obtain that the corresponding densities converge to the density of the standard normal.

The proof is algebraically much nicer than the general case involving moment generating functions because we have such a nice closed form expression for the moment generating function of exponential random variables.

Problem 2. We have

$$\widehat{f}(y) \ = \ \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx.$$

Recall that a function f is Schwartz if for any non-negative integers m and n there is a constant $C_{m,n}$ such that $|f^{(n)}(x)| \le C_{m,n}/(1+x^2)^m$. In other words, f and all its derivatives decay faster than the reciprocal of any polynomial. We write the denominator as $1+x^2$ and not x^2 or x so that our bounds are well-defined when $x=0$ too.

We 'evaluate' the integral by integrating by parts. There are no convergence issues, as f and all its derivatives are of rapid decay. We set $u = f(x)$, $dv = e^{-2\pi ixy}dx$ so $du = f'(x)dx$ and $v = (-2\pi iy)^{-1}e^{-2\pi ixy}$. *If you are uncomfortable integrating functions such as* $e^{-2\pi ixy}$*, break it up into* $\cos(2\pi xy) + i\sin(2\pi xy)$. As f is Schwartz, the boundary terms (evaluating $u(x)v(x)$ at $\pm\infty$) vanish because of the rapid decay of f, and thus

$$\widehat{f}(y) \ = \ \frac{1}{2\pi iy}\int_{-\infty}^{\infty} f'(x)e^{-2\pi ixy}dx,$$

where there is no minus sign as the minus sign from the derivative of the exponential factor cancels the minus sign from integrating by parts. Repeating this $k-1$ times yields

$$\widehat{f}(y) \ = \ \frac{1}{(2\pi iy)^k}\int_{-\infty}^{\infty} f^{(k)}(x)e^{-2\pi ixy}dx,$$

where $f^{(k)}$ represents the k^{th} derivative of f. As f is Schwartz, $f^{(k)}(x)$ decays faster than any polynomial in x, and hence the integral above exists. It is therefore bounded independent of y (so long as $y \neq 0$, but we may assume $y \neq 0$ as we care about the behavior for large y), say by C_k for some C_k depending on $f^{(k)}$.

Explicitly, we use the fact that the absolute value of an integral is at most the integral of the absolute value, and then use $|e^{i\theta}| = 1$ for real θ. *Again, if you are not comfortable with working with complex valued functions, write it as a cosine piece plus a sign piece, and work with each piece individually.* We have

$$\left|\widehat{f}(y)\right| \ \le \ \frac{1}{|2\pi y|^k}\int_{-\infty}^{\infty} |f^{(k)}(x)|dx.$$

As $|f^{(k)}(x)| \le C_{k,1}/(1+x^2)$, the integral converges and is finite (it actually equals π). Letting $C_k = \pi C_{k,1}/(2\pi)^k$, we find

$$\left|\widehat{f}(y)\right| \ \le \ \frac{C_k}{|y|^k},$$

which gives the desired decay.

The purpose of this problem is to give the beginning of a proof of an important fact, namely that the Fourier transform of a Schwartz function is a Schwartz function. This is an ingredient in the proof of the Fourier Inversion Theorem. This is the first, and most important, step in proving this claim. We now need to show all the derivatives of $\widehat{f}$ have the appropriate decay. We sketch the proof. It all starts with the relation

$$\widehat{f}(y) \ = \ \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx.$$

We differentiate with respect to y a total of n times. As f is Schwartz, we may interchange the differentiation with respect to y and the integration, and we find

$$\widehat{f}^{(n)}(y) \ = \ \int_{-\infty}^{\infty} f(x)\cdot(-2\pi ix)^n e^{-2\pi ixy}dx \ = \ (-2\pi i)^n \int_{-\infty}^{\infty} g(x)e^{-2\pi ixy}dx,$$

where

$$g(x) \ = \ f(x)x^n.$$

The proof is completed by noting that $g(x)$ is a Schwartz function and then using the bounds from the first part of the problem. The arguments here illustrate another important property of the Fourier transform: there is a relation between multiplying our original function by x^n and taking n derivatives of the Fourier transform. This interplay is one reason why the Fourier Transform is extremely useful in solving differential equations.

Problem 3. For (a), note

$$\mu'_{2k} \ = \ \int_{-B}^{B} x^{2k}f(x)dx \ \le \ B^{2k}\int_{-B}^{B} f(x)dx \ = \ B^{2k};$$

this is because x^{2k} is largest when $x = \pm B$, and thus the $2k^{\text{th}}$ moment is at most equal to what we obtain when we replace x^{2k} with B^{2k}. In attacking problems like this, it helps to think about what we should aim for. It is very good if we end up integrating $f(x)$, as we know this integrates to 1. How large can the factor of x^{2k} be? As $-B \le x \le B$, $x^{2k} \le B^{2k}$, and we do not decrease (and almost surely increase) the value of the integral if we replace x^{2k} with B^{2k}.

For (b), the claim is true. This problem seems like a converse to (a). In part (a) we showed that if the function is supported in $[-B, B]$ then the $2k^{\text{th}}$ moment is at most B^{2k}; here we are saying we know the limit of the $1/2k$ root of μ'_{2k} is at most B – we want to show that this implies f lives in $[-B, B]$.

We give a proof by contradiction. Assume not. Then there is a positive probability of our random variable X with density f taking on a value x with $|x| > B$. We thus find that for some $\epsilon > 0$ there is a $p > 0$ such that $\text{Prob}(|X| \ge B + \epsilon) > p$. We need to pass from this to a lower bound for μ'_{2k} which is so large that the $1/2k$ power tends to infinity, as this will imply that there is no finite B such that all of the probability lives in $[-B, B]$. We need to think of how to incorporate this into a lower bound. A little thought tells us that if we want a lower bound, and we know there is a positive probability of $|X| \ge C > B$, it might be worth restricting the integral defining μ'_{2k} to $|x| \ge C$. If we do this, we'll be able to replace all x^{2k} in this region with C^{2k} and get a lower bound.

In particular, this implies

$$\mu'_{2k} \ = \ \int_{-\infty}^{\infty} x^{2k}f(x)dx \ > \ (B+\epsilon)^{2k}\text{Prob}(|X| \ge B+\epsilon) \ = \ p\cdot(B+\epsilon)^{2k}.$$

Taking the $1/2k$ root yields

$$\mu_{2k}'^{1/2k} \ > \ p^{1/2k}(B+\epsilon).$$

As $p > 0$, as $k \to \infty$ we have $p^{1/2k} \to 1$, and thus

$$\lim_{k\to\infty} \mu_{2k}'^{1/2k} \ \geq \ B+\epsilon$$

(note the greater than becomes a greater than or equal to in the limit). This contradicts our assumption, and thus X is supported in $[-B, B]$. The importance of this problem is that an understanding of the even moments gives enormous amounts of information about the probability density; in particular, if the moments don't grow too rapidly than the density is supported in some finite interval, while if they grow very rapidly then the density is not supported in any finite interval. Why are we looking at the even moments to the exclusion of the odd moments? The problem with the odd moments is that they can be small even if the function is of unbounded support (just think of the standard normal).

For (c), our random variable is not supported in any finite interval $[-B, B]$. To prove this, it suffices to study the lower bound, as this will imply the moments grow so rapidly that the density is not finitely supported. Clearly $(2k)!! > k!$; both have the same number of terms, but each factor of $(2k)!! = 2k\cdot(2k-2)\cdots 2$ is twice the corresponding factor of $k! = k\cdot(k-1)\cdots 1$. By Stirling's formula, for large k we have $k! \sim k^k e^{-k}\sqrt{2\pi k}$. Thus

$$\mu_{2k}'^{1/2k} \ \geq \ k!^{1/2k} \ \sim \ (k/e)^{1/2}(2\pi k)^{1/2k} \ \geq \ \sqrt{k/2}.$$

From the first part of the problem, if f is supported in $[-B, B]$ then $\mu_{2k}'^{1/2k} \ \leq \ B$; as $k \to \infty$ we see there is no fixed B that can work, and thus our function has unbounded support.

There are lots of other relations we could have used, such as $(2k)!! \geq \sqrt{(2k)!}$ and $(2k)!! = 2^k k!$.

These three problems are meant to give a bit of the flavor of the theory of moments. In particular, knowing a little bit about how large (or small) the moments are in the limit allows us to deduce some things about the distribution, such as whether or not it lives in some finite interval.

Problem 4. The argument given is wrong. The problem is that while the main terms are correct in each case, the secondary terms are ignored. We are in a situation where the two main terms cancel, and thus it is essential that we understand these secondary terms as well. One way to do this problem is to take the exponential of the logarithm. This is a generalization of multiplying by 1, as we do nothing but in a useful way.

Let

$$P_N \ = \ \left(1+\frac{x}{N}\right)^{N^2} \cdot \left(1-\frac{x}{N}\right)^{N^2}.$$

Taking such an approach, we find we must study, as $N \to \infty$,

$$\log P_N \ = \ N^2 \log\left(1 + \frac{x}{N}\right) + N^2 \log\left(1 - \frac{x}{N}\right).$$

Taylor expanding, using

$$\begin{aligned} \log(1+u) &= u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots \\ \log(1-u) &= -u - \frac{u^2}{2} - \frac{u^3}{3} - \cdots \end{aligned}$$

we find

$$\begin{aligned} \log P_N &= N^2\left(\frac{x}{N} - \frac{x^2}{2N^2} + O(N^{-3})\right) + N^2\left(-\frac{x}{N} - \frac{x^2}{2N^2} + O(N^{-3})\right) \\ &= -x^2 + O(N^{-1}), \end{aligned}$$

which implies

$$P_N \ = \ e^{-x^2} e^{-O(1/N)},$$

and thus

$$\lim_{N\to\infty} P_N \ = \ e^{-x^2}.$$

There is a simpler way to see this. Note

$$\begin{aligned} \lim_{N\to\infty} P_N &= \lim_{N\to\infty} \left[\left(1 + \frac{x}{N}\right)^{N^2} \cdot \left(1 - \frac{x}{N}\right)^{N^2}\right] \\ &= \lim_{N\to\infty} \left[\left(1 + \frac{x}{N}\right) \cdot \left(1 - \frac{x}{N}\right)\right]^{N^2} \\ &= \lim_{N\to\infty} \left(1 - \frac{x^2}{N^2}\right)^{N^2}; \end{aligned}$$

note, however, that this limit is just the definition of e^{-x^2}. The only difference between the above and the standard definition is that we are using N^2 instead of N; however, if $N \to \infty$ then $N^2 \to \infty$ as well.

For this problem, it is much easier to use the second approach. The idea is that the algebra is nicer here because we have $1 + u$ times $1 - u$, which is just $1 - u^2$ (for $u = x/N$). This factorization and reinforcement occurs in many problems. We present both proofs as, for general questions, we don't have the luxury of exploiting such a nice factorization. Pattern recognition is extremely important; we can easily stare for hours at something that we know without realizing we know it because it is presented in a different light. After thinking to combine the two factors, we then have to see that $N \to \infty$ is the same as $N^2 \to \infty$, and what we have is just the definition of $\exp(-x^2)$.

It's worth emphasizing that this technique of combining factors surfaces all the time in advanced number theory courses.

It's worth spending a few moments thinking about the false argument given above. There, we have a limit of a product We'd like to say that this equals the product of the limit, but we must be careful. Those two limits are ∞ and 0, and 0 times ∞ is not defined. It is clearly illegal to do what we did, namely to say the limit of the product is the product of the limits, but we don't take the limits, instead letting $N \to \infty$ in some parts but not in other places so as to end up with e^{xN} and e^{-xN}.

In general, depending on how things tend to infinity and zero, we can make $\infty \cdot 0$ equal almost anything. Consider

$$\begin{aligned} \lim_{N\to\infty} N^2 \cdot \frac{1}{N} &= \infty \\ \lim_{N\to\infty} N^2 \cdot \frac{\pi}{N^2} &= \pi \\ \lim_{N\to\infty} N \cdot \frac{1}{N^3} &= 0. \end{aligned}$$

If we were to say the limit of a product is the product of the limits, in each case we get $\infty \cdot 0$, and yet each case has a different answer.

Perhaps my favorite problem involving $\infty \cdot 0$ argues that (at least in this case!) it should be -1. Here's the example: consider the product of the slopes of any two perpendicular lines not parallel to the coordinate axes. It's a very nice exercise to prove this product is -1. As it equals -1 so long as the lines are not parallel to the coordinate exes, it is natural to define the product of the slopes here (0 for the x-axis and ∞ for the y-axis) to be -1.

Problem 5. We first note that $|A+A|$ and $|A-A|$ are not changed by mapping each $x \in A$ to $\alpha x + \beta$ for any fixed α and β. The effect of this transformation is to take all the sums and multiply by α and then add 2β, while for the differences it multiplies by α. For definiteness, imagine we have the mapping $x \mapsto 3x+5$ and take $a_1 = 4$, $a_2 = 7$. Then originally the sum is 11 and the difference is -3. Our elements map to 17 and 26, and now the sum is $43 = 3 \cdot 11 + 2 \cdot 5$ and the difference is $-9 = 3 \cdot (-3)$.

As we have an arithmetic progression, we use the following map: $x \mapsto (x - n_0)/d$. This maps our initial arithmetic progression to the new progression $\{0, 1, 2, \dots, n\}$. It is very easy to compute the sumset and difference set here. The set of sums is just $\{0, 1, \dots, 2n\}$ and the set of differences is just $\{-n, \dots, n\}$. As both of these sets have $2n + 1$ elements, the sumset and difference set are the same size.

Perhaps an easier way to view the problem is the following. Clearly $A = \{0, 1, \dots, n\}$ lead to $A+A = \{0, 1, \dots, 2n\}$ and $A-A = \{-n \dots, n\}$. To see this, note the smallest element of $A + A$ is clearly obtained from $0 + 0$, while the largest comes from $n + n$. A little inspection shows we can get everything in between. Instead of going from our original arithmetic progression to this, we go the other way (start here and end up with the original progression). *This is an example of the very useful method of reverse engineering.* We first change variables by mapping $\{0, 1, \dots, n\}$ to $\{0, d, \dots, nd\}$. This mapping changes all sums and all differences by a factor of d, but does not alter how many such factors there are. Similarly, if we include the translation n_0, then all sums increase by the same amount (namely $2n_0$) while the differences are unchanged.

Learning what change of variables to do and when to do it is a terrific skill that frequently takes a lot of practice, but it is important. For example, frequently difficult integrals are in integral table books, but in an equivalent manner that can be missed unless we see the 'clever' change of variables problems.

This is a very important problem in additive number theory. The reason is that such sets are balanced (i.e., they have as many sums as differences). These sets are often used as a starting point in the construction of sum-dominated sets in the following way: if we start with an arithmetic progression, the hope is that by tweaking it slightly we can add one more sum than difference, and thus end up with a sum-dominated (or more sum than difference) set.

12. HW #12

Due Thursday December 10 (though you may place in my mailbox up till 10am on Friday 12/11):

DO ANY THREE OF THE PROBLEMS BELOW, BUT ONE OF THE THREE PROBLEMS MUST BE PROBLEM #1. *If you choose to do either problem 5 or 6 (you of course may elect to do both), you must email me your .tex file and .pdf, and note on the homework you submit to the grader which of these problems you elected to do.*

(1) **Everyone must do this one:** Take two homework or exam problems where you lost points this semester because your logic was incorrect (i.e., what you wrote was wrong and not just you left the problem blank). Write a short TeX document where you state the problem and explain your reasoning as to why you made the mistake you did, and email me the TeX file. Make sure you give the file a name which begins with your lastname (this makes it easy for me to keep track of who's work I'm viewing).

(2) Come to my office and give a 5 to 10 minute presentation on some topic on probability that we have not covered in class. This could be the solution to a problem we haven't done from a section we've studied, summarizing a section we haven't studied, summarizing a paper, ... The point of this exercise is to get practice orally presenting information (in addition to being good in its own right, it helps if you ever need a letter of recommendation, as I can then talk about your presentation skills).

(3) Consider a random variable X with the standard Laplace distribution, so the density is $f(x) = \exp(-|x|)/2$. According to Chebyshev's inequality or theorem, what is the probability X is more than k standard deviations from the mean? Do you think this a good estimate? What is the actual probability?

(4) Consider a Cauchy random variable X, so $f(x) = \frac{1}{\pi(1+x^2)}$. What does Chebyshev's theorem or inequality say about the probability $|X| > 2009$? Estimate

this probability.

(5) Write up a problem of your choosing and a solution.

(6) Read a paper involving probability and give a one page summary.

(7) Work on one of the research projects mentioned in class. If you elect to do this, contact me.

Problem 3. For the Laplace distribution, the density is $f(x) = \exp(-|x|)/2$. The mean is clearly zero, and thus the variance is

$$\mathbb{E}[X^2] \ = \ \int_{-\infty}^{\infty} x^2 e^{-|x|}\frac{dx}{2} \ = \ \int_0^{\infty} x^2 e^{-x} dx.$$

One way to do this would be to integrate by parts twice. For the first we set $u = x^2$ and $dv = e^{-x}$, and we see $du = 2xdx$ and $v = -e^{-x}$. Continuing in this manner yields $\mathbb{E}[X^2] = 2$, or the variance is 2. A faster approach is to recall the definition and properties of the Gamma function:

$$\Gamma(s) \ = \ \int_0^{\infty} e^{-s}x^{s-1}dx \text{ (if } \mathfrak{Re}(s) > 0), \quad \Gamma(n+1) \ = \ n! \text{ (if } n \text{ is a positive integer).}$$

Our integral expression for $\mathbb{E}[X^2]$ is just $\Gamma(3) = \Gamma(2+1)$, and thus the answer is just $2! = 2$.

The first part of the problem asks for the estimate from Chebyshev's theorem for being more than k standard deviations from the mean. Chebyshev's theorem (also known as Chebyshev's inequality) states

$$\text{Prob}(|X - \mu| \ge k\sigma) \ \le \ \frac{1}{k^2}$$

for any random variable X with finite mean μ and finite variance σ^2. Thus Chebyshev estimates this probability as $1/k^2$ (which of course is useless for $k \le 1$).

Chebyshev's theorem holds for all densities with finite second moment; it thus uses very little information about the distribution itself. It must hold for both uniform random variables and exponential and Gaussians. It is thus quite likely that this greatly overestimates the true probability. What is the true probability? As the variance is 2, the standard deviation is $\sqrt{2}$ and we are reduced to computing

$$\text{Prob}(|X| \ge k\sqrt{2}) \ = \ \int_{|x|\ge k\sqrt{2}} e^{-|x|}\frac{dx}{2} \ = \ \int_{k\sqrt{2}}^{\infty} e^{-x}dx \ = \ e^{-k\sqrt{2}} \ = \ \left(\frac{1}{e^{\sqrt{2}}}\right)^k.$$

Thus the true answer is *significantly* smaller than Chebyshev. For example, when $k = 2$ Chebyshev gives .25 while the true answer is about .06; for $k = 10$ Chebyshev gives .01 versus about $7 \cdot 10^{-7}$, which is magnitudes smaller! To really drive home the point, if $k = 100$ then Chebyshev gives .0001 while in fact the probability is less than $4 \cdot 10^{-62}$!

Problem 4. As this problem is similar to the previous, see Problem #3 for a statement of Chebyshev's theorem. As the Cauchy distribution has infinite variance, Chebyshev's theorem has *nothing* to say about the probability of $|X| \ge 2009$. The density function of the standard Cauchy distribution is $f(x) = \frac{1}{\pi(1+x^2)}$, with cumulative distribution function $F(x) = \frac{\arctan(x)}{\pi}$. Thus

$$\begin{aligned}\text{Prob}(|X| \geq 2009) &= \int_{|x|\geq 2009} \frac{dx}{\pi(1+x^2)}\\ &= \frac{2}{\pi}\int_{2009}^{\infty} \frac{dx}{1+x^2}\\ &= \frac{2}{\pi}\left(\arctan(\infty) - \arctan(2009)\right)\\ &= \frac{2}{\pi}\left(\frac{\pi}{2} - \arctan(2009)\right).\end{aligned}$$

How should we approximate $\arctan(2009)$? Plugging into Mathematica gives

$$\arctan(2009) \approx 1.5702985667563642558, \quad \frac{\pi}{2} \approx 1.5707963267948966192.$$

This is fine if we have a computer at our disposal; however, what if we don't?

All is not lost, as we know the power of Taylor series, so let's simple expand $\arctan(x)$ and then evaluate it at $x = 2009$. The problem with this is that we would need a Taylor series expansion about infinity, not 0, as we want to see what happens when we evaluate the probability for large x. We can rectify this by recalling

$$\arctan(x) = \frac{\pi}{2} - \arctan\left(\frac{1}{x}\right) \quad \text{if } x > 0.$$

The Taylor series expansion for $\arctan$ is

$$\arctan(u) = \sum_{n=0}^{\infty} \frac{(-1)^n u^{2n+1}}{2n+1} = u - \frac{u^3}{3} + \frac{u^5}{5} - \frac{u^7}{7} + \cdots.$$

Thus

$$\arctan\left(\frac{1}{2009}\right) \approx \frac{1}{2009},$$

with an error on the order of $1/2009^3$. Substituting gives

$$\arctan(2009) = \frac{\pi}{2} - \frac{1}{2009},$$

so

$$\text{Prob}(|X| \geq 2009) = \frac{2}{\pi}\left(\frac{\pi}{2} - \arctan(2009)\right) \approx \frac{2}{2009\pi};$$

this gives a predicted value of 0.000316884, the actual answer of 0.000316883882... is quite close (and we could obtain better approximations with more terms).

Another way to approximate the answer is to say that, since $x \geq 2009$, $1/(1+x^2) \approx 1/x^2$; the error in such an approximation is $1/x^2(1+x^2)$, which when integrated over $[2009, \infty)$ is of size $1/2009^3$. Using this approximation, we get

$$\text{Prob}(|X| \geq 2009) \approx \frac{2}{\pi}\int_{2009}^{\infty} \frac{dx}{x^2} = \frac{2}{2009\pi},$$

the same answer we found with Taylor series and arctan identities (but significantly less work!).

For comparison's sake, the probability that $|X| \geq 2009$ for the standard normal is less than $10^{-876423}$, while for the Laplace distribution it is about $1.265 \cdot 10^{-1234}$.

In the original proof we used an identity for $\arctan(x)$, namely $\arctan(1/x) = \pi/2 - \arctan(x)$ for $x > 0$. There are many ways to prove this. One way is to use the fact that $\arctan(x)$ is the anti-derivative of $1/(1+x^2)$. As $\arctan(0) = 1$ and $\arctan(\infty) = \pi/2$, we have

$$\arctan\left(\frac{1}{x}\right) \quad = \quad \frac{1}{\pi}\int_0^{1/x} \frac{dt}{1+t^2}.$$

We change variables, setting $u = 1/t$ so $du = -dt/t^2$ or $dt = -du/u^2$. The region of integration is now from ∞ to x, which becomes x to ∞ as we have a negative sign. Thus

$$\begin{aligned}
\arctan\left(\frac{1}{x}\right) &= \int_\infty^x \frac{1}{1+(1/u^2)} \frac{-du}{u^2} \\
&= \int_x^\infty \frac{du}{1+u^2} \\
&= \arctan(\infty) - \arctan(x) \\
&= \frac{\pi}{2} - \arctan(x)
\end{aligned}$$

as claimed. A faster proof is to note that for any angle $\theta \in (0, \pi/2)$, $\tan(\theta)$ and $\tan(\frac{\pi}{2} - \theta)$ have reciprocal tangents; thus $\arctan(x) + \arctan(1/x) = \pi/2$.

We end with one more way to solve the problem. Either changing variables or using our arctan identity, we see the problem is equivalent to evaluating

$$\int_0^{1/2009} \frac{dt}{1+t^2}.$$

We use the geometric series to expand $(1+x^2)^{-1}$, interchange the integral and the sum, and then evaluate term by term. Thus we have

$$\begin{aligned}
\arctan(x) &= \int_0^x \frac{dt}{1+t^2} \\
&= \int_0^x \frac{dt}{1-(-t^2)} \\
&= \int_0^x \sum_{n=0}^\infty (-t^2)^n dt \\
&= \sum_{n=0}^\infty \frac{(-1)^n t^{2n+1}}{2n+1},
\end{aligned}$$

which is the claimed Taylor series expansion for arctan; while some work is needed to justify the arguments above, this is much faster than computing the Taylor expansion of

$\arctan$ from first principles.

Finally, since $\frac{d}{dx}\arctan(x) = 1/(1+x^2)$ is so essential for this problem, we quickly review it's proof. Using the quotient rule, we know $\frac{d}{dx}\tan(x) = 1/\cos^2(x)$. We use a very useful identity: if f and g are inverse, differentiable functions such that $f(g(x)) = x$, then $f'(g(x))g'(x) = 1$ or $g'(x) = 1/f'(g(x))$. Letting f denote the tangent function and g the arctangent function, we have

$$\frac{d\arctan(x)}{dx} \ = \ \frac{1}{\tan'(\arctan(x))} \ = \ \cos^2(\arctan(x)).$$

While this is a solution, as written $\cos(\arctan(x))$ is not that illuminating. We now show that $\cos(\arctan(x)) = 1/\sqrt{1+x^2}$. To see this, let $\theta = \arctan(x)$, so $\tan(\theta) = x$. We construct a right triangle with side adjacent to the angle θ equal to 1 and side opposite the angle θ equal to x. Clearly this will have $\tan(\theta) = x/1 = x$ as desired. Further, by Pythagoras the hypotenuse's length is $\sqrt{1^2+x^2} = \sqrt{1+x^2}$. Then $\cos(\theta) = 1/\sqrt{1+x^2}$; however, $\theta = \arctan(x)$, so $\cos(\arctan(x)) = 1/\sqrt{1+x^2}$ as claimed.

In summary, this problem reviews many of the key properties of the Cauchy distribution. Even though this density has infinite second moments, there is a nice, closed form expression for the cumulative distribution function, and thus it is easy to integrate. We have simple series expansions for the cumulative distribution function, and can compute the relevant probabilities without too much difficulty. The hardest part is knowing which identities to use to simplify the algebra or to approximate the answer.