Doob's Optional Stopping Theorem

The Doob's optional stopping time theorem is contained in many basic texts on probability and Martingales. (See, for example, Theorem 10.10 of *Probability with Martingales*, by David Williams, 1991.) The essential content of the theorem is that you can't make money (in expectation) by buying and selling an asset whose price is a martingale. Precisely, the theorem states that if you buy the asset at some time and adopt any strategy at all for deciding when to sell it, then the expected price at the time you sell is the price you originally paid. Thus—if market price is a martingale—you cannot make money in expectation by "timing the market."

Let be the probability space. Let T be a map from to the set of positive integers. We think of $T(\omega)$ as giving the time at which the asset will be sold if the price sequence is $S(0), S(1), S(2), \ldots$ (where each $S(i)$ is a random variable, i.e., a real-valued function of ω). We say that T is a **stopping time** if the event that $T(\omega) = n$ depends only on the values $S(i)$ for $i \leq n$. In other words, the decision of whether to sell the stock at time n depends only on the history of the stock up until time n, and not on the future values of the stock (which the investor hasn't seen yet).

Doob's Optional Stopping Theorem: If the sequence $S(0), S(1), S(2), \ldots$ is a bounded martingale, and T is a stopping time, then the expected value of $S(T)$ is $S(0)$.

Most real world asset prices are not martingales, even in theory. So why is this theorem relevant to finance?

1. Many asset prices behave approximately like martingales in the short term.

2. According to the **fundamental theorem of asset pricing**, as presented in Zastawniak and Capiński (see the text for the precise conditions of the theorem), the discounted price $\frac{S(n)}{A(n)}$, where A is a risk-free asset, *is* a martingale with respected to the **risk neutral probability**.

3. Sequences of conditional expectations of a quantity—involving conditioning on increasing amounts of information—are martingales. For example, let C be the amount of oil available for drilling under a particular piece of land. Suppose that ten geological tests are done

that will ultimately determine the value of C. Let C_n be the **conditional expectation** of C *given* the outcome of the first n of these tests. Then the sequence $C_0, C_1, C_2, \ldots, C_{10} = C$ is a martingale.

SOME MARTINGALE PROBLEMS:

1. Suppose Harriet has 7 dollars. Her plan is to make one dollar bets on fair coin tosses until her wealth reaches either 0 or 50, and then to go home. What is the expected amount of money that Harriet will have when she goes home? What is the probability that she will have 50 when she goes home?

2. Consider a contract that at time N will be worth either 100 or 0. Let $S(n)$ be its price at time $0 \leq n \leq N$. If $S(n)$ is a martingale, and $S(0) = 47$, then what is the probability that the contract will be worth 100 at time N?

3. Pedro plans to buy the contract in the previous problem at time 0 and sell it the first time T at which the price goes above 55 or below 15. What is the expected value of $S(T)$?

4. Suppose $S(N)$ is with probability one either 100 or 0 and that $S(0) = 50$. Suppose further there is at least a sixy percent probability that the price will at some point dip to below 40 and then subsequently rise to above 60 before time N. Prove that $S(n)$ cannot be a martingale.

18.600 Probability and Random Variables

Fall 2019

Permutations, Poker, and Powerball

18.600 Problem Set 1

Welcome to your first 18.600 problem set! There will be ten problem sets this semester, each with a different theme, and each including a mix of problems of my own design and problems from the Sheldon Ross 8th edition textbook. Before we begin the problems, let me provide some basic information and experience-based advice to help you get more out of the course as a whole.

1. Recognize that, like courses in the 18.0x series, this is a fundamental course, meant to be accessible to students from every department at MIT. On the other hand (and here it is crucial to set expectations correctly) the course is *very challenging* for many students, and in some ways more advanced than the 18.0x courses. You should not be surprised or disappointed when you don't know how to do the problems right away, or when you need to consult a second or third source to understand a concept.

2. Attend lectures! In principle, it may be possible to learn much of the course material by clicking through lecture slides in Adobe fullscreen mode on your bedroom laptop, but this will not necessarily make you happy. People who come to lecture and stay engaged (ask questions, answer questions, etc.) learn more quickly, have more fun, and remember longer. If the lectures seem a little fast, try reading the textbook and/or slides in advance and come prepared to ask questions. If they seem a little slow (or if the topic is one you have seen before) try just showing up and letting the lecture be your first exposure. The fraction of students who have seen the lecture material before will be significant during the first few weeks, but will decline very quickly after that.

3. Use the textbook. Part of the reason for including problems from the textbook is to remind you that the textbook exists, and to encourage you to read it (any of the 6th through 10th editions will do). All of the course material (except for the parts about martingales and Black-Scholes, which will be covered in a separate handout) is covered in the textbook. There are many inexpensive ways to get electronic or hard copies of the textbook online (check out ebay, amazon, google, etc.)

4. Start the problem sets early and come to office hours if you have questions. The problem sets are more challenging than the exams and they serve a different purpose. They are meant to be educational in their own right. If you are one of the lucky few for whom the problems are easy, you will still learn a lot by thinking through the concepts and applications. You are free to collaborate with other students, look up material on the internet or in books, offer each other hints (though

not full solutions or answers) on Piazza, and ask me and the TAs for ideas during office hours and recitations. Note however that some of the problems are reused from prior years, and you are definitely *not* allowed to access or consult prior year problem set solutions. (Prior year *exam* solutions, on the other hand, are posted on the public course webpage, and you are welcome to use those.)

5. If you are doing well in the course, try to help out by answering (as well as asking) questions on Piazza. This will help solidify your own understanding and will be appreciated by fellow students (as well as your TAs and me). I am going to try to restrain myself from answering most basic math questions on Piazza (so students have more of a chance to answer questions for each other) but if you post a question on Piazza and it remains unanswered after 48 hours, let me know by email and I'll look into it. If you have a personal problem or a complaint or a request (e.g., "Could you use a different chalk color?"), you should email the TAs or me directly, rather than posting something on Piazza, which will be more of a public forum.

6. Spare at least a *little* time for thinking and exploring that has nothing to do with your grade. Like looking up and reading a bit more about mathematical or practical issues raised in problem sets. Ponder some big questions about applications. What are we doing wrong in medicine? In traffic management? In college admissions? In teaching? In food preparation? Is there simple advice that, if followed, would make us all better off? Is there other commonly accepted advice that we should all stop following? How can we find out what these things are? How can probability help? As the course progresses, we will see many problems with applications; but each problem is the beginning of a conversation, not the end.

Plagiarism policy: We will abide by the MIT plagiarism policy (google *MIT plagiarism*) if it appears that a student has lifted a problem set solution from an online source (such as a manual or a previous year solution set) without attribution.

Now the problems. The first problem set is about basic combinatorics. Cards, hats, permutations, balls, binomial and multinomial coefficients. It will help to keep these stories in mind as the course progresses.

A. FROM ROSS 8th EDITION CHAPTER ONE:

1. **Problem 10:** In how many ways can 8 people be seated in a row if

(a) there are no restrictions on the seating arrangement?

(b) persons A and B must sit next to each other?

(c) there are 4 women and 4 men and no 2 men or 2 women can sit next to each other?

(d) there are 5 men and they must sit next to each other?

(e) there are 4 married couples and each couple must sit together?

2. **Theoretical Exercise 11:** The following identity is known as Fermat's combinatorial identity:

$$\binom{n}{k} = \sum_{i=k}^{n} \binom{i-1}{k-1} \quad n \geq k.$$

Give a combinatorial argument (no computations are needed) to establish this identity. *Hint:* Consider the set of numbers 1 through n. How many subsets of size k have i as their highest-numbered member?

B. Consider permutations $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. There are $n!$ such permutations altogether. Of these permulations...

1. How many have only one cycle, i.e., have the property that $\sigma(1), \sigma \circ \sigma(1), \sigma \circ \sigma \circ \sigma(1), \ldots$ cycles through all elements of $\{1, 2, \ldots, n\}$?

2. How many have exactly two cycles, one of length k (where $1 \leq k \leq n-1$) and one of length $n-k$?

3. How many are involutions, i.e., have the property that for each j we have $\sigma \circ \sigma(j) = j$? (*Hint:* Argue that if σ is an involution then each j is either a fixed point — i.e., satisfies $\sigma(j) = j$ — or part of a cycle of length two. Compute the number of involutions with exactly k cycles of length 2, and then write your overall answer as a sum over k.)

C. In a standard deck of 52 cards, there are 4 suits, and 13 cards of each suit. Using such a deck, there are $\binom{52}{13}$ ways to form a bridge hand containing 13 cards. How many of these hands have the property that:

1. All 13 cards belong to the same suit.

2. Exactly 2 of the 4 suits are represented in the hand.

3. Exactly 3 of the 4 suits are represented in the hand.

4. All 4 of the suits are represented (i.e., there is at least one card of each suit).

There is a hint on the next page, but don't look before you need to.

Hint: This one is legitimately tricky. If it helps, you can try generalizing the problem. Imagine that instead of 4 suits you have a deck with k suits, 13 cards of each suit—and then let $N_k(m)$ be the number of ways to produce a 13 card hand *from this deck* that has exactly m suits represented. Maybe you can build a table containing all the values of $N_k(m)$ for $k \in \{1, 2, 3, 4\}$ and $1 \leq m \leq k$. Can you show that $\sum_{m=1}^{k} N_k(m) = \binom{13k}{13}$? Can you show that $N_k(m) = \binom{k}{m} N_m(m)$? Does this help you complete the table?

D. In the US lottery game of Powerball one is required to choose an (unordered) collection of five numbers from the set $\{1, 2, \ldots, 69\}$ (the white balls) along with another number from the set $\{1, 2, \ldots, 26\}$ (the red ball). So there are $\binom{69}{5} \cdot 26 = 292201338$ possible Powerball outcomes. You make your selection (five white, one red), the Powerball people choose theirs randomly, and you win if there is a match. Suppose that you have already chosen your numbers (the unordered set of five white, and the one red). How many possible Powerball outcomes match *exactly one* of your five white numbers (regardless of whether they match the red number)? How many match exactly two of your five white numbers? How many match exactly three? How many match one red ball *plus* exactly two white balls? Now, divide each of these numbers by 292201338 to produce a *probability* of seeing that outcome and use a calculator to give a numerical value. Write a sentence about what seems interesting or surprising about these values.

Remark: People in the US spend over 70 billion per year on lottery tickets (about 32 percent of which is returned in big lottery payouts). That's over two hundred dollars per person, with many players spending thousands of dollars a year. The psychological appeal may be hard for us to understand. But the fact that regular players get "close" now and then (matching two or three numbers) may be part of what keeps them coming back. If you are one of the *many* people who buys more than 1000 lottery tickets per year, you will probably match four of the six balls at some point during your life. If you have a dozen friends who do the same thing, one of them will probably match five of six balls at some point, which will *seem* very close. You can double check your computations by looking up the odds on the Powerball wikipedia page.

E. Derive the following formulas, which will be useful later in this course:

1. **Normal density formula:** $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 1$. (Multiply both sides by $\sqrt{2\pi}$ and square both sides to get

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy = 2\pi.$$

   Then derive this by computing the integral in polar coordinates. You can look up the derivation in the book if you get stuck.)

2. **Poisson mass formula:** $\sum_{k=0}^{\infty} e^{-\lambda}\lambda^k/k! = 1$ if $\lambda > 0$. Hint: recall (or look up) the Taylor expansion for the function $f(\lambda) = e^{\lambda}$.

3. **Binomial sum formula:** $\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = 1$ where $p + q = 1$ and n is a positive integer. Hint: try expanding $(p + q)^n$.

4. **Factorial formula:** $\int_0^{\infty} x^n e^{-x} dx = n!$. (Assume $n \geq 0$ is an integer and use integration by parts and induction.)

Store these formulas in long term memory and write "Got it!"

Remark: You will at some point have to learn a few formulas for this class: in particular, those that appear in red on the so-called story sheet posted on the exam page. But it will turn out that a surprising number of them (perhaps a majority) are obtained in some way from the four formulas listed above. Internalizing these few facts now will help you a lot going forward. These formulas (or close variants) are among those appearing in garageband clip about identities I posted last year at `http://math.mit.edu/~sheffield/2018600/kindofthing.mp4`. I am not sure about its musical merits, but the clip does contain some nice identities, as well as some problems that you will see later in this course.

18.600 Probability and Random Variables

Fall 2019

Axioms and assumptions

18.600 Problem Set 2

Welcome to your second 18.600 problem set! We will continue to explore some combinatorics along with probability and the axioms of probability.

Before we get to work, let's indulge in a bit of reflection. When we say "The probability that A will happen is p" where does p come from? Sometimes the evidence convinces pretty much everyone that A will or will not happen. Informally, the probability that a predicted lunar eclipse will happen on schedule is pretty much 1, and the probability that Mars and Jupiter will collide this month is pretty much 0. In other simple situations (die rolls, coin tosses, etc.) experience may lead us to agree on probabilities that aren't 0 or 1. The assumption that all outcomes are equally likely (for random permutations or die rolls or coin tosses) is sometimes a natural starting point. This assumption is implicitly made in a few of the problems here.

In more complicated real world settings, one can sometimes define the *risk neutral* probability, a probability measure derived from the market prices of contracts whose values depend on future events. If we want to know the risk neutral probability that a given candidate will win an election, or that an athletic team will win a game, we can look at betting markets. (Check out electionbettingodds.com, predictit.com, oddschecker.com, betfair.com, and similar websites.) As we will see later in the course, if we want to know the risk neutral probability that the price of a share of Apple stock will exceed some value by the end of the year, we can work this out by looking at current prices of *derivatives* (contracts whose future value depends on future share prices). The total amount of money at stake in derivative markets is estimated at over a quadrillion dollars per year (try googling derivatives quadrillion).

Some argue that betting markets set up perverse incentives. If I buy a contract that gives me $\$500,000$ if my house burns down, that's useful insurance. But if I buy a contract that gives me $\$500,000$ if *your* house burns down, that gives me an unhealthy incentive to burn your house down. People similarly worry about a world in which hedge funds can bet that a company will collapse and then actively cause it to collapse. Rules are required to prevent such things, but foolproof (and evil-genius-proof) rules are hard to design and enforce.

On the other hand, one might argue that the absence of betting markets is part of the reason that some questions in politics and law are divisive. It is hard to place a bet on the proposition that "my candidate would do more to advance long term happiness and prosperity than yours" or "my client is innocent," so there is no market mechanism for producing a commonly accepted probability. Different groups can *claim* to have different probability estimates, the expression of which may advance their own agendas, but without a market we cannot tell which parties would actually be *willing to bet money* at the corresponding rates. Some studies claim that people answering questions about the economy are both more accurate and less partisan when they are paid (even a very small amount) for correct answers. Maybe there is something to be said for having money on the line.

A. FROM ROSS 8TH EDITION CHAPTER TWO:

1. **Problem 25:** A pair of dice is rolled until a sum of either 5 or 7 appears. Find the probability that a 5 occurs first. *Hint.* Let E_n denote the event that a 5 occurs on the nth roll and no 5 or 7 occurs on the first $(n-1)$ rolls. Compute $P(E_n)$ and argue that $\sum_{n=1}^{\infty} P(E_n)$ is the desired probability.
2. **Theoretical Exercise 10:** Prove that $P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E^c FG) - P(EF^c G) - P(EFG^c) - 2P(EFG)$.
3. **Theoretical Exercise 20:** Consider an experiment whose sample space consists of a countably infinite number of points. Show that not all points can be equally likely. Can all points have a positive probability of occurring?

B. On Sep. 12, 2019, I looked up Democratic presidential nomination contracts on two trading platforms: predictit and betfair. According the table below, one could purchase a predictit YES contract on Elizabeth Warren for 35 cents. This is worth \$1 if Warren wins the nomination and \$0 otherwise. Similarly, one could purchase a predictit NO contract on Elizabeth Warren for 66 cents. This is worth \$1 if Warren is *not* elected and \$0 otherwise. (When you purchase a NO contract for 66 cents you are technically *selling* a YES contract for 34 cents to somebody else offering to buy it for that price. The gap between the 34 offer-to-buy and the 35 offer-to-sell is called the *bid-ask spread*. Because predictit only trades in integers, the spread is typically 1 cent — which is why YES and NO prices sum to 101 instead of 100.)

	predictit YES	predictit NO	betfair YES	betfair NO
Elizabeth Warren	35	66	34.5	66.2
Joe Biden	27	74	21.7	79.2
Bernie Sanders	17	84	13.5	87.2
Andrew Yang	12	89	6.1	94.6
Kamala Harris	11	90	10.9	89.6
Pete Buttigieg	9	92	5.0	95.8
Hillary Clinton	5	96	3.6	96.7
Cory Booker	4	97	2.5	98.3
Beto O'Rourke	3	98	0.8	99.6
Julian Castro	2	99	1.0	99.1
Tulsi Gabbard	3	98	1.0	99.3
Tom Steyer	3	98	0.8	99.9
Amy Klobuchar	2	99	0.7	99.5

Because the betting probabilities are not the same on the two sites, there may be opportunities for *arbitrage*, i.e., opportunities to make money without taking risk. (For now, let us ignore taxes, fees, interest and the risk associated with either of these companies going

bankrupt before the election; of course these things matter in practice.) If I buy Biden NO on predictit for 74 and Biden YES on betfair for 21.7 then I have only spent 95.7. But I am guaranteed to win 100. If I purchase all 13 predictit NO contracts then the price comes to $11.80. But because at least 12 of these candidates are guaranteed to lose, I will be guaranteed to win at least $12 before fees (and about $11.88 after fees). Note that per predictit rules, I only have to leave enough money on the site to cover the amount I will owe in the worst case scenario; this means that I can take my 8 cents of profit now *without* having to leave any money on the platform for the duration between now and the election. So: 8 free cents.

1. Using the chart above, find three other arbitrage opportunities, i.e., ways to make a risk free profit (again ignoring the caveats mentioned above: fees, tax, interest, etc.)

2. Per the efficient market hypothesis, arbitrage opportunities should not exist. Give a short speculative explanation for why there seem to be arbitrage opportunities here. (This will not be graded except to check that you thought about it.)

REMARK: Predictit traders are only allowed to spend $850 on each candidate. (This was part of the deal making predictit legal in the US, unlike betfair.) So one can buy 42500 shares of Klobuchar YES (2 cents each) but only 858 shares of Klobuchar NO (99 cents each). Some people think these constraints cause unlikely YES contracts to be overpriced. In theory, arbitrageurs should correct this overpricing by buying lots of NO contracts (thus bidding down the YES prices); but if each potential arbitrageur can only make a *little* "free money" this way (due to spending caps), it may just not be worth the effort.

REMARK: Suppose that you think the "true" value of Warren YES is 34.5 but have a personal reason for wanting to own a Warren YES share. Then there are two things you can do. One, you can go straight to the site and buy a share for 35 (basically overpaying by half a cent). Or two, you can put up an "offer" to buy for 34 and wait until somebody is willing to take the other side of the bet (by buying NO at 66). If you take the second approach, you might get a better price. But you will have to wait in line behind everyone else offering to buy at 34. And there is some risk that you never get to the front of the line (perhaps the price of Warren YES shares goes up before you manage to buy at 34) — and the scenarios in which you fail to get to the front of the line might be precisely the scenarios in which you would most like to own Warren YES. Sometimes on predictit the lines can be long (tens or even hundreds of thousands of shares) and because trades can only take place at integer values, you cannot jump to the front of the line by offering 34.1. For example, last I checked there are over 350,000 offers to *sell* Mark Zuckerberg YES for 1 cent. (Zuckerberg is not running.) Some people making those offers would be willing to sell for half a cent or a tenth of a cent. But because of the strict integer rule, they can't make those offers. They have to wait in line.

C. Suppose that there are $M \geq 1$ job candidates and $N \geq 1$ companies with job openings. Each candidate (independently, uniformly at random) develops a serious interest in one of the

N companies and each company (independently, uniformly at random) develops a serious interest in one of the M candidates. What is the probability that there is at least one company-candidate pair that are seriously interested in each other? (Hint: let E_j be the event that the jth applicant's interest is requited. Use inclusion-exclusion on these events. You can write the probability as a sum. Don't worry about simplifying further.) Later in this course (after we discuss *additivity of expectation*) we will find that the *expected* (a.k.a. *average*) number of candidates with requited interest is $\sum_{j=1}^{M} P(E_j)$. Compute this quantity as a function of M and N, and write a sentence about whether you consider the answer surprising.

D. Alice and Bob are playing a game of tennis and have reached the game state called "deuce." From here the players keep playing points until one player's point-win total exceeds the other player's total by 2, at which point the player ahead by 2 is declared winner of the game. Suppose that Alice wins each point with probability p (independently of all previous points) and Bob wins each point with probability $q = (1 - p)$ (independently of all previous points). Find the probability that Alice wins the game, as a function of p. (Hint: consider what happens over the course of the next *two* points. Either Alice wins both and the game is over, or Bob wins both and the game is over, or each player wins a point and the players are back where they started. Compute the probabilities of these three outcomes. Then apply the ideas from the first problem on this problem set.) Based on your answer, do you agree or disagree with the following statement? *If Alice is k times as likely as Bob to win a point, then Alice is k^2 times as likely as Bob to win the game if the current score is deuce.*

E. The online comic strip xkcd.com has a "random" button one can click to choose one of the previous $n \approx 2200$ strips. Assume there are exactly 2200 numbered strips and that clicking the "random" button yields the kth strip, where k is chosen uniformly from $\{1, 2, \ldots, n\}$. If one observes m strips this way, what is the probability that one sees at least one strip more than once? (This is a variant of the birthday problem.) Give rough numerical values for $m = 36$ and $m = 56$ and $m = 78$. (Hint: try going to wolframalpha.com and entering something like Prod[(1-k/2200), {k,0,35}].) Based on your answer, do you agree or disagree with the following statement? *The number of clicks required before you see the same strip twice is a random quantity whose median is about* 56*, and which lies between* 36 *and* 78 *about half the time*

F. A "gender reveal" party is held to announce the gender of an expected newborn. 15 cups are filled in advance with colored beads and covered: if the baby is a boy, 8 are filled with blue beads and 7 with pink. If the baby is a girl, 7 are filled with blue and 8 with pink. When the audience arrives the cups are knocked over (revealing bead colors) in a uniformly random order until the audience has seen 8 cups of the same color (and thereby knows the gender). Let N be the number of cups turned over before the gender is known. Compute the probability that $N = k$ for $k \in \{8, 9, 10, \ldots, 15\}$. Is the probability that $N = 15$ (and one has to wait to the very end) more or less than 1/2? (This problem was communicated to me by a friend who found $N = 15$ in a real life implementation and wanted to know how surprising that was.)

G. The following is a popular and rather instructive puzzle. A standard deck of 52 cards (26 red and 26 black) is shuffled so that all orderings are equally likely. We then play the following game: I place the deck face down and begin turning over the cards from the top of the deck one at a time so that you can see them. At some point (before I have turned over all 52 cards) you say "now." At this point I turn over the next card and if the card is red, you receive one dollar; otherwise you receive nothing. You would like to design a strategy to maximize the probability that you will receive the dollar. How should you decide when to say "now"?

Your first observation is that a good time to say "now" is when you know a high fraction of the cards remaining in the deck are red. On the other hand, if you wait for this fraction to increase, there is a chance you'll see more red cards while you wait, so that the fraction *decreases*. What's the right way to balance these concerns, i.e., what is the *optimal* strategy for deciding when to say "now"? The next page has a hint but don't look until you have to.

HINT: Imagine a variant of the game in which, after you say "now," I turn over the *bottom* card on the deck. Observe that in this variant, it makes no difference when you say "now" (since you're going to see the same bottom card of the deck regardless). Now try to argue that your probability of winning with a given strategy in the modified game is the same as your probability of winning with that strategy in the original game. Conclude that in the original game, it also makes no difference which strategy you choose. Your odds of winning the dollar are .5 regardless.

REMARK: The fraction of red cards in the deck turns out to be something called a *martingale* and the fact that it makes no difference when you bet can be derived from something called the *optional stopping theorem*. We'll have more on this at the end of the course. But I like this puzzle because it's something you can appreciate right now. This is a strategy game that Warren Buffett and a chimpanzee would play equally well. Unlike the game shown here `https://www.youtube.com/watch?v=JkNV0rSndJ0` and playable here `http://www.softschools.com/games/memory_games/ayumus_game/`

18.600 Probability and Random Variables

Fall 2019

Conditional probability
18.600 Problem Set 3

Welcome to your third 18.600 problem set! Conditional probability is defined by $P(A|B) = P(AB)/P(B)$, which implies

$$P(B)P(A|B) = P(AB) = P(A)P(B|A),$$

and dividing both sides by $P(B)$ gives Bayes' rule:

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)},$$

which we may view as either a boring tautology or (after spending a few hours online reading about Bayesian epistemology, Bayesian statistics, etc.) the universal recipe for revising a worldview in response to new information. Bayes' rule relates $P(A)$ (our Bayesian *prior*) to $P(A|B)$ (our Bayesian *posterior* for A, once B is given). If we embrace the idea that our brains have subjective probabilities for *everything* (existence of aliens, next year's interest rates, Sunday's football scores) we can imagine that our minds continually use Bayes' rule to update these numbers. Or least that they would if we were clever enough to process all the data coming our way.

By way of illustration, here's a fanciful example. Imagine that in a certain world, a *normal* person says 10^5 things per year, each of which has a 10^{-5} chance (independently of all others) of being truly horrible. A *truly horrible* person says 10^5 things, each of which has a 10^{-2} chance (independently of all others) of being truly horrible. Ten percent of the people in this world are truly horrible. Suppose we meet someone on the bus and the first thing that person says is truly horrible. Using Bayes' rule, we conclude that this is probably a truly horrible person.

Then we turn on cable news and see an unfamiliar politician saying something truly horrible. Now we're less confident. We don't know how the quote was selected. Perhaps the politician has made 10^5 recorded statements and we are seeing the only truly horrible one. So we make the quote selection mechanism part of our sample space and do a more complex calculation.

The problem of selectively released information appears in many contexts. For example, lawyers select evidence to influence how judges and jurors calculate conditional probability *given* that evidence. If I'm trying to convince you that a number you don't know (but which I know to be 49) is prime, I could give you some selective information about the number without telling you exactly what it is (it's a positive integer, not a multiple of 2 or 3 or 5, less than 50) and if you don't consider my motives, you'll say "It's probably prime."

Note also that legal systems around the world designate various "burdens of proof" including *probable cause*, *reasonable suspicion*, *reasonable doubt*, *beyond a shadow of a doubt*, *clear and convincing evidence*, *some credible evidence*, and *reasonable to believe*. Usually, these terms lack clear meaning as numerical probabilities (does "beyond reasonable doubt" mean with probability at least .95, or at least .99, or something else?) but there is an exception: *preponderance of evidence* generally indicates that a probability is greater than fifty percent, so that something can be said to be "more likely than not."

An interesting question (which I am not qualified to answer) is whether numerical probabilities should be assigned to the other terms as well.

A. FROM TEXTBOOK CHAPTER THREE:

1. Problem 47: An urn contains 5 white and 10 black balls. A fair die is rolled and that number of balls is randomly chosen from the urn. What is the probability that all of the balls selected are white? What is the conditional probability that the die landed on 3 if all the balls selected are white?

B. A medical practice uses a "rapid influenza diagonistic test" to get a quick (under 30 minute) assessment of whether a patient has the flu. The **sensitivity** of the test (i.e., the fraction of the time it returns a positive result if the patient has the disease) is .5 while the **specificity** (i.e., the fraction of the time it returns a negative result if the person does not have the flu) is .9. In other words, people without the flu are relatively unlikely (10 percent chance) to get a false positive, but people *with* the flu have a larger chance (50 percent) to get a false negative (e.g., because the particular strain of flu isn't picked up by the test, or virus somehow didn't make it onto the swab).

Suppose that based on time of year and symptoms (fever, chills, cough, etc.) the doctor thinks *a priori* that the event F that a patient has the flu with probability $P(F) = .7$. Assume further that the doctor believes that the specificity/sensitivity results mentioned above apply to this individual (given what is known), so that if T is the event that the test comes back positive, we have $P(T|F) = .5$ and $P(T|F^c) = .1$. After the doctor administers the test and discovers that the test is negative, what is the doctor's *a posteriori* estimate of the probability that the patient has the flu? In other words, what is $P(F|T^c)$? What is $P(F|T)$? Give approximate percentages.

Remark: Google *RIDT specificity and sensitivity* to see actual estimates of these values (which vary with the study, the type of test, the flu strains prevalent in a given year, etc.) One practical decision a doctor might make is whether to prescribe an antiviral medication (like Tamiflu) that is thought to reduce symptom duration by about one day on average *if* a person has the flu, and zero days otherwise. (By comparison, a flu vaccine that reduces the risk of a 10-symptom-day flu during a season from 20 percent to 10 percent would also decrease the expected number of symptom days by one. Google *Tamiflu effectivness* and *flu vaccine effectiveness* for actual data on such things, which apparently vary quite a lot from year to year and place to place.) One might guess that if a person has a p chance of having the flu, then taking the drug decreases the expected number of symptom days by p (one day on average if flu is really there, zero otherwise). If p is below some threshold the doctor and patient may conclude that the cost (time it takes to fill prescription, price of drug, small side effect risk) just isn't worth it. (Another hard-to-measure consideration: how much do vaccines/antivirals decrease the risk of infecting others?)

C. Suppose that a fair coin is tossed infinitely many times, independently. Let X_i denote the outcome of the ith coin toss (an element of $\{H, T\}$). Compute:

1. the conditional probability that exactly 7 of the first ten tosses are heads *given* that exactly 9 of the first 20 tosses are heads.

2. the probability that there exists an infinite arithmetic progression such that $X_i = H$ for all i in that arithmetic progression. In other words, there exist positive integers a and b such $X_i = H$ whenever $i \in \{a, a+b, a+2b, a+3b, a+4b, \ldots\}$. (Hint: use the countably additivity axiom.)

3. the probability that the pattern HHTTHTTHT appears at least once in the sequence $X_1, X_2, X_3, \ldots$

4. the probability that *every* finite-length pattern appears *infinitely many times* in the sequence $X_1, X_2, X_3, \ldots$

D. On Interrogation Planet, there are 730 suspects, and it is known that exactly one of them is guilty of a crime. It is also known that any time you ask a guilty person a question, that person will give a "suspicious-sounding" answer with probability .9 and a "normal-sounding" answer with probability .1. Similarly, any time you ask an innocent person a question, that person will give a suspicious-sounding answer with probability .1 and a normal-sounding answer with probability .9. (And these probabilities apply *regardless* of how the suspect has answered questions in the past; in other words, once a person's guilt or innocence is fixed, that person's answers are *independent* from one question to the next.)

Interrogators pick a suspect at random (all 730 people being equally likely) and ask that person nine questions. The first three answers sound normal but the next six answers all sound suspicious. The interrogators say "Wow, six suspicious answers in a row. Only a one in a million chance we'd see that from an innocent person. This person is obviously guilty." But you want to do some more thinking. Given the answers thus far, compute the conditional probability that the suspect is guilty. Give an exact numerical answer.

E. Suppose that the quantities $P[A|X_1], P[A|X_2], \ldots, P[A|X_k]$ are all equal. Check that $P[X_i|A]$ is proportional to $P[X_i]$. In other words, check that the ratio $P[X_i|A]/P[X_i]$ does not depend on i. (This requires no assumptions about whether the X_i are mutually exclusive.)

Remark: This can be viewed as a mathematical version of Occam's razor. We view A as an "observed" event and each X_i as an event that might "explain" A. What we showed is that if each X_i "explains" A equally well (i.e., $P(A|X_i)$ doesn't depend on i) then the conditional probability of X_i *given* A is proportional to how likely X_i was a *a priori*. For example, suppose A is the event that there are certain noises in my attice, X_1 is the event that there are squirrels there, and X_2 is the event that there are noisy ghosts. I might say that $P(X_1|A) >> P(X_2|A)$ because $P(X_1) >> P(X_2)$. Note that after looking up online definitions of "Occam's razor" you might conclude that it refers to the above tautology *plus* the common sense rule of thumb that $P(X_1) > P(X_2)$ when X_1 is "simpler" than X_2 or "requires fewer assumptions."

F. On Cautious Science Planet, science is done as follows. First, a team of wise and well informed experts concocts a hypothesis. Experience suggests the hypotheses produced this way are correct ninety percent of the time, so we write $P(H) = .9$ where H is the event that the hypothesis is true. Before releasing these hypotheses to the public, scientists do an additional experimental test (such as a clinical trial or a lab study). They decide in advance what constitutes a "positive" outcome to the experiment. Let T be the event that the positive outcome occurs. The test is constructed so that $P(T|H) = .95$ but $P(T|H^c) = .05$. The result is only announced to the public if the test is positive. (Sometimes the test involves checking whether an empirically observed quantity is "statistically significant." The quantity $P(T|H)$ is sometimes called the *power* of the test.)

(a) Compute $P(H|T)$. This tells us what fraction of published findings we expect to be correct.

(b) On Cautious Science Planet, results have to be replicated before they are used in practice. If the first test is positive, a second test is done. Write $\tilde{T}$ for the event that the second test is positive, and assume the second test is like the first test, so that $P(\tilde{T}|HT) = .95$ but $P(\tilde{T}|H^cT) = .05$. Compute the reproducibility rate $P(\tilde{T}|T)$.

(c) Compute $P(H|T\tilde{T})$. This tells us how reliable the replicated results are. (Pretty reliable, it turns out—your answer should be close to 1.)

On Speculative Science Planet, science is done as follows. First creative experts think of a hypothesis that would be rather surprising and interesting if true. These hypotheses are correct only five percent of the time, so we write $P(H) = .05$. Then they conduct a test. This time $P(T|H) = .8$ (lower power) but again $P(T|H^c) = .05$. Using these new parameters:

(d) Compute $P(H|T)$.

(e) Compute the reproducibility rate $P(\tilde{T}|T)$. Assume the second test is like the first test, so that $P(\tilde{T}|HT) = .8$ but $P(\tilde{T}|H^cT) = .05$.

Remark: If you google Nosek reproducibility you can learn about one attempt to systematically reproduce 100 psychology studies, which succeeded a bit less than 40 perent of the time. Note that $P(\tilde{T}|T) \approx .4$ is (for better or worse) closer to Speculative Science Planet than Cautious Science Planet. The possibility that $P(H|T) < 1/2$ for real world science was famously discussed in a paper called *Why Most Published Research Findings Are False* by Ioannidis in 2005. A more recent mass replication attempt (involving just *Science* and *Nature*) allowed scientists to bet on whether a study would be replicated and found that to some extent scientists were good at predicting such things. See `https://www.nature.com/articles/d41586-018-06075-z`.

Questions for thought: What are the pros and cons of the two planets? Is it necessarily bad for $P(\tilde{T}|T)$ and $P(H|T)$ to be low in some contexts (assuming that people know this and don't put too much trust in single studies)? Do we need to do larger and more careful studies? What improvements

can be made in fields like medicine, where controlled clinical data is sparse and expensive but life and death decisions have to be made nonetheless? And I do mean expensive. The cost of recruiting and pre-screening a *single* Alzheimer's patient for trial is $100,000, per this article
`https://www.nytimes.com/2018/07/23/health/alzheimers-treatments-trials.html`
These questions go well beyond the scope of this course, but we will say a bit more about the tradeoffs involved when we study the central limit theorem.

G. **Doomsday:** Many people think it is likely that intelligent alien civilizations exist *somewhere* (though perhaps so far separated from us in space in time that we will never encounter them). When a species becomes roughly as advanced and intelligent as our own, how long does it typically survive before extinction? A few thousand years? A few millions years? A few billion years? Closely related question: how many members of such a species typically get to exist before it goes extinct?

Let's consider a related problem. Suppose that one factory has produced a million baseball cards in 10,000 batches of 100. Each batch is numbered from 1 to 100. Another factory has produced a million baseball card in 1,000 batches of 1,000, each batch numbered from 1 to 1,000. A third factory produced a million baseball card in 100 batches of 10,000, with each batch numbered from one to 10,000. You chance upon a baseball card from one of these three factories, and *a priori* you think it is equally likely to come from each of the three factories. Then you notice that the number on it is 74.

(a) Given the number you have seen, what is the conditional probability that the card comes from the first factory? The second? The third?

Now consider the following as a variant of the card problem. Suppose that one universe contains 10^{50} intelligent beings, grouped into civilizations of size 10^{12} each. Another universe contains 10^{50} intelligent beings, grouped into civiliations of size 10^{15} each. A final universe contains 10^{50} intelligent beings, grouped into civilizations of size 10^{18} each. You pick a random one of these 3×10^{50} beings and learn that before this being was born, exactly $141,452,234,521$ other beings were born in its civilization.

(b) What is the conditional probability that the being comes from the first universe?

Remark: The *doomsday argument* (google it) is that it is relatively likely that human civilization will disappear within thousands of years — as opposed to lasting millions of years — for the following reason: *if* advanced civilizations typically lasted for millions of years (with perhaps 10 billion beings born per century), then it would seem *coincidental* for us to find ourselves among the first few thousand. People disagree on what to make of this argument (what the Bayesian prior on civilization length should be, what to do with all the other information we have about our world, what measure to put on the set of alternative universes, etc.) Maybe the argument at least makes people think about the *possibility* of near-term human extinction, and whether preparing for apocalyptic scenarios (giant asteroids, incurable plagues, nuclear war, climate disaster, supervolcanos, resource depletion, the next ice age, etc.) might improve our chance of surviving a few thousand (or million or billion) more years.

18.600 Probability and Random Variables

Fall 2019

Expectation, covariance, binomial, Poisson

18.600 Problem Set 4

Welcome to your fourth 18.600 problem set! The interesting topics we have discussed in lecture include the linearity of expectation, the bilinearity of covariance, and the notion of *utility* as used in economics. (Under certain "rationality" assumptions everyone has a utility function whose expectation they seek to maximize.) We will see in this problem set how these ideas play a role in some important (though perhaps overly simplistic) theories from finance (MPT and CAPM). We will have more on binomial/Poisson random variables and a chance to learn about Siegel's paradox.

A. FROM TEXTBOOK CHAPTER FOUR:

1. Problem 23: You have \$1000, and a certain commodity presently sells \$2 per ounce. Suppose that after one week the commodity will sell for either \$1 or \$4 an ounce, with these two possibilities being equally likely.

   (a) If your objective is to maximize the expected amount of money that you possess at the end of the week, what strategy should you employ?

   (b) If your objective is to maximize the expected amount of the commodity that you possess at the end of the week, what strategy should you employ?

Remark: Look up Siegel's paradox. It's pretty interesting.
`http://mindyourdecisions.com/blog/2012/03/15/siegels-paradox-about-exchange-rates/`

B. On ACT Planet, Jack is preparing to take a test called the Math ACT. Jack knows his stuff but is error prone under pressure, and because of this he only gets the right answer 85 percent of the time. His success probability is the same for all problems (no matter how hard they are for others) and his outcomes are independent from one problem to another. If he gets his expected 51 out of 60 answers correct, it comes to a Math ACT score of 30. (On ACT planet, the score conversion table is the same each time the test is given.)

Jack wants to attend "Thirty Four or Higher University" (known as TFOHU) for which he needs a 34 on the Math ACT, which requires at least 55 out of 60 correct. Fortunately, the university only requires that he obtain that score once. He is not required to report scores that fall below the treshold. Jack decides to invest the time and money to take the exam 12 times. Assuming Jack's abilities remain constant, what is the probability that he gets a sufficiently high score (i.e., at least 55 correct answers) at least once? Give both an algebraic expression and an approximate percentage. (A calculator like `http://stattrek.com/online-calculator/binomial.aspx` may help you compute the numerical value.)

C. Suppose that during each minute there is a 1 in $500,000$ probability that there is an accident at a particular intersection (independently of all other minutes). Using the approximation of $500,000$

minutes per year, we expect to see 1 accident per year on average. One year somebody proposes to install a new kind of stoplight to reduce accidents. You believe *a priori* that there is a 1/3 chance that the new stoplight is *effective*, in which case it will reduce the accident rate by fifty percent, and a 2/3 chance it will have no effect. The new stoplight is installed and during the next year there are no accidents. Using Poisson approximations, compute your updated estimate of the probability that the light is effective.

Remark: It is often hard to tell whether preventative measures against rare events are having an effect. With twenty years of data we might be more confident, but by that point accident rates may have changed for other reasons (e.g., self driving cars). On the other hand the $k!$ in the Poisson denominator means that *large* numbers are *extremely* unlikely. If we suddenly see 10 accidents in one year, we should seriously question our assumption that the number is Poisson with $\lambda = 1$ or $\lambda = 1/2$.

D. Larry the Very Subprime Lender gives loans of size \$10,000. In 25 percent of cases, the borrower pays back the loan quickly with no interest or fees. In 50 percent of cases, the borrower disappears (moves away, declares bankruptcy, dies) without paying anything. In 25 percent of cases, the borrower pays back the loan slowly and — after years of ballooning interest payments, hefty fees, etc. — pays Larry a total of \$100,000. However, in this scenario, Larry has to give \$60,000 to third parties (repo services, foreclosure lawyers, eviction teams, bill collectors, etc.) in order to get the borrower to pay the \$100,000. Compute the following:

(a) The expectation and variance of the *net* amount of profit Larry makes from each loan (after subtracting collection expenses and the initial \$10,000 outlay).

(b) The expectation and variance of the *net* amount a given borrower ends up paying (i.e., amount paid minus amount borrowed).

Note: You might have ethical concerns with Larry's (unrealistic) business model. Google *payday loan* or *buy here pay here* or (for purely negative side) *predatory lending* (maybe start with the Wikpedia articles) if you want more realistic information about the legal and moral issues involved in lending to populations with high default risk. It is a large and complicated subject.

E. Define the covariance $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$. Define the *corelation* between X and Y to be $\mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$. One can show that the correlation between any two random variables is always between -1 and 1. *Very* roughly speaking, the correlation is high (i.e., close to 1) if X tends to be high when Y is high and tends to be low when Y is low.

1. Check that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$, that $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$, and that $\mathrm{Cov}(\cdot,\cdot)$ is a bilinear function of its arguments. That is, if one fixes one argument then it is a linear function of the other. For example, if we fix the second argument then for real constants a and b we have $\mathrm{Cov}(aX + bY, Z) = a\mathrm{Cov}(X,Z) + b\mathrm{Cov}(Y,Z)$.

2. If $\mathrm{Cov}(X_i, X_j) = ij$, find $\mathrm{Cov}(X_1 - X_2, X_3 - 2X_4)$.

3. If $\mathrm{Cov}(X_i, X_j) = ij$, find $\mathrm{Var}(X_1 + 2X_2 + 3X_3)$.

4. Suppose that V and $X_1, X_2, \ldots, X_n$ are random variables, that $\mathrm{Var}(V) = 1$, that $\mathrm{Cov}(X_i, V) = b_i$ for each i and that $\mathrm{Cov}(X_i, X_j) = c_{i,j}$ for each pair i, j, where b_i (for $1 \leq i \leq n$) and $c_{i,j}$ (for $1 \leq i, j \leq n$) are known constants. Suppose for some fixed constants $a_1, a_2, \ldots, a_n$, we write $X = \sum a_i X_i$. Then demonstrate the following:

   (a) $\mathrm{Cov}(X, V) = \sum_{i=1}^{n} a_i b_i$

   (b) $\mathrm{Var}(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i c_{i,j} a_j$.

   (c) The correlation between X and V is $\frac{\sum a_i b_i}{\sqrt{\sum a_i c_{i,j} a_j}}$.

   With some calculus and linear algebra (which I won't make you do) you can use the above to find a choice of $a_1, a_2, \ldots, a_n$ that *maximizes* the correlation between X and V.

Remark: Many university rating systems (and clickbaity lists like "top ten cities for singles" or "best companies to work for") are constructed using a weighted sum X of measurements X_i each believed to be *correlated* with some (hard to define) *overall* value V. For example, US News measures what fraction of alumni donate, how much professors are paid, what fraction of faculty have PhDs, what fraction of students were top ten percent in high school, etc. In each case, the measured quantity is not something students necessarily care about *for its own sake* — rather, it is believed to be *correlated* with things they care about. The QS World University Rankings (where MIT is first) use a weighted sum of six different quantities (citations per faculty, academic and employer reputation, etc.)

Many feel that these rankings are useful in holding universities accountable and conveying at least a rough sense of where the strong universities are. Others are more critical. One problem is that the rankings depend heavily on the choice of measured quantities (the X_i) and the weights (the a_i). These choices often seem arbitrary and *ad hoc*, and differ greatly from one ranking system to another. Another problem is that even if we pretend there is a quantity V that represents *overall value*, and even if we have defined an X such that the correlation between X and V is high (say .8) across all universities, it is not clear that the correlation remains high if we restrict attention to, say, the top 20 universities. (Maybe a statistic like "5 $*$ *height in inches* minus 2 $*$ *age in years*" is well correlated with basketball ability in a randomly chosen adult, but not in a randomly chosen NBA player.) A final concern is that institutions may "game the system" by taking actions that increase X without increasing V. Some worry that these actions waste resources, and may even decrease V. (The adage that "When a measure becomes a target, it ceases to be a good measure" is sometimes called Goodhart's law.) Google *us news rankings controversy* for some seriously anti-ranking polemics.

F. Use the following identities (some more well known than others) to solve the problems below:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \qquad \sum_{n=0}^{\infty} \frac{1}{n!} = e \qquad \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} = \log(2) \qquad \sum_{n=2}^{\infty} \binom{n}{2}^{-1} = 2$$

$$\sum_{n=1}^{\infty} n^s = \prod_{p \text{ prime}} (1 - p^s)^{-1} \text{ if } s < -1 \qquad \frac{1}{\pi} = \frac{\sqrt{8}}{99^2} \sum_{k=0}^{\infty} \frac{(4k)!(1103 + 26390k)}{(k!)^4 396^{4k}}$$

1. A town launches an annual marathon. Let M_n and W_n denote the winning times for men and women during the nth year. Assume the M_n are independently chosen from some continuous probability measure and the W_n are independently chosen from another continuous probability measure (the precise choice doesn't matter). Observe that the probability that there is a women's record during the nth year (i.e., that $W_n < W_k$ for all $k < n$) is $1/n$.

   (a) Compute the expected number of years during which there is simultaneously a men's record and a women's record (assuming the annual marathon will continue forever).

   (b) Compute the probability that there is *never* simultaneously a men's record and women's record during a *prime* numbered year.

   (c) Let K be the number of the first year during which there is *not* a women's record. (So $K \geq 2$.) Compute the expectation of K. You can use the fact that $E[K] = \sum_{k=0}^{\infty} P[K > k]$.

   (d) Let L be number of year when the first woman winner first sees her record beaten. Compute the probability L is even. Hint: observe that $P(L \text{ even}) = P(L > 1) - P(L > 2) + P(L > 3) - P(L > 4) + \ldots$.

   (e) Compute the expected number of years $n \geq 2$ when the most recent two winning men are fastest, i.e. $\max\{M_n, M_{n-1}\} < M_k$ for $k < n - 1$.

2. In a future unified muggle/wizard sorting process, an infinite line of people is waiting to be sorted. At each step, the sorting hat independently declares a person to be a wizard with probability 1/99, in which case it sorts the individual into one of four houses (each with probability 1/4). In between sortings, if it is ever determined that

   (a) everyone sorted thus far has been a wizard, and

   (b) each of the four wizard houses has an equal number of people,

   then this rare occurrence is celebrated in the obvious way: namely, the sorting is paused, an interhouse Quidditch tournament is conducted, and each person in the winning house receives 26390 Galleons, while the house itself receives a trophy worth 1103 Galleons. (A trophy is also awarded during the "empty" game that occurs at the beginning when all houses have zero people.) Compute the expectation of the total value that will ever be awarded (trophies included).

Remark: These problems are from a garageband clip about identities I posted last year `http://math.mit.edu/~sheffield/2018600/kindofthing.mp4`. (Most mathematical music videos don't contain much math; this is an exception, for better or worse.) You can use the answers at the end of the clip to check your numbers, but you need to show work to get credit (i.e., at least write in a few words what the individual terms/factors represent). If you think events like $E_p = \{\text{no record occurs in } p\text{th year}\}$ are *independent* you should say why. **Hint:** Fix n and let σ be the permutation such that year j was $\sigma(j)$th fastest year for the women's time (among the first n years). Start by arguing by symmetry that all such permutations are equally likely.

G. Instead of maximizing her expected wealth $E[W]$, Jill maximizes $E[U(W)]$ where $U(x) = -(x - x_0)^2$ and x_0 is a large positive number. That is, Jill has a *quadratic utility function*. (It may seem odd that Jill's utility declines with wealth once wealth exceeds x_0. Let us assume x_0 is large enough so that this is unlikely.) Jill currently has W_0 dollars. You propose to sample a random variable X (with mean μ and variance σ^2) and to give her X dollars (she will lose money if X is negative) so that her new wealth becomes $W = W_0 + X$.

1. Show that $E[U(W)]$ depends on μ and σ^2 (but not on any other information about the probability distribution of X) and compute $E[U(W)]$ as a function of x_0, W_0, μ, σ^2.

2. Show that given μ, Jill would prefer for σ^2 to be as small as possible. (One sometimes refers to σ as *risk* and says that Jill is *risk averse*.)

3. Suppose that $X = \sum_{i=1}^n a_i X_i$ where a_i are fixed constants and the X_i are random variables with $E[X_i] = \mu_i$ and $\text{Cov}[X_i, X_j] = \sigma_{ij}$. Show that in this case $E[U(W)]$ depends only on the μ_i and the σ_{ij} (but not on any other information about the joint probability distributions of the X_i) and compute $E[U(W)]$. Hint: first compute the mean and variance of X.

Remark: We conclude (assuming quadratic utility) that portfolio builders care *only* about expectations and covariances of items in their portfolio. This idea underlies the (1990 Nobel Prize Winning) *Modern Portfolio Theory* (MPT) and *Capital Asset Pricing Model* (CAPM). Before these theories, it was believed that when the *variance* of an asset return is high, the *expected* return should be higher as well (the *risk premium*) because otherwise people wouldn't buy risky assets. MPT and CAPM predict that one gets a risk premium for *systemic risk* (the part of the variance explained by correlation with the *market portfolio*, defined to be the sum total of all risky assets) but not for *idiosyncratic risk* (exposure to which can be reduced by diversification). These theories also predict that everyone's optimal investment strategy is to put some (investor-dependent) fraction of their money in a risk-free asset and the remainder in the market portfolio (which we think of as a giant index fund). Google MPT and CAPM to read about how well or poorly these theories match reality.

4. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with the same mean and variance. Show that among all random variables of the form $\sum_{i=1}^n a_i X_i$ (where the a_i are non-negative numbers with $\sum_{i=1}^n a_i = W$ for some fixed constant W) the one with the smallest variance is the one with $a_i = W/n$ for each i.

Remark: If the (presumed i.i.d.) X_i are returns on n investments, then the above implies that one minimizes risk by dividing wealth equally among the investments. In the story below, the X_i are the overall stock market returns in different *years*. Suppose you plan to contribute K dollars to your child's college fund annually for 18 years, dividing wealth between a (zero interest) safe investment and a (risky) stock index fund. You decide in advance that on the ith year, you will invest a_i dollars in stocks. Then (if $\sum a_i$ is held constant) your variance is minimized if you invest the same amount in stocks each year. If you instead keep a fixed *percentage* of wealth in stocks each year, then your final value will depend most heavily on market performance during the later years (when you have the most money in the account). This is why some financial planners recommend being more *aggressive* (i.e., open to higher risk in exchange for higher expectation) during early years and more *conservative* as you get closer to using the money. *However,* if the money were all contributed *upfront* with no annual K-dollar influx (and you assumed *logarithmic* utility) you could make a case for keeping a fixed *percentage* in stocks. More on this later.

Remark: People often say utility functions should be strictly concave (negative second derivative) to explain risk aversion... but is that necessarily true? Here is a naive story about charitable giving. Suppose your utility function is given by your own health/comfort plus a constant c times the sum of the health/comfort of all other humans on the planet. For example, if $c = .01$, then you are mostly selfish, but you would be willing to give up a comfort for yourself if it would enable more than 100 strangers to enjoy the same comfort. You'd give up your life if you could save more than 100 other lives. Utilitarians might theorize that it is a good to thing that $c > 0$ (so that we help others when we can make a big difference) but maybe also a good thing that $c < 1$ (since a little selfishness might be efficient in practice). As you acquire more money, there may come some point at which you believe that the marginal value of another dollar to you (in added health/comfort) is *less than* c times the amount a dollar donated to a global charity with relatively high expected impact (like those profiled at givewell.org) would increase health/comfort for others. After that point, in principle you should donate *all* of your additional money to charity. If this is indeed your plan, then your utility function might be very close to linear for a long time after that point, since the amount of good you do in a huge global effort is roughly linear in the amount you give.

Remark: Some economists say that in reality charitable giving should be modeled as a consumptive good (that happens to have positive externality — google *"warm glow giving"*) that has to compete with other consumptive goods among even the very wealthy. This point of view might predict actual behavior better than what I sketched above.

Remark: A "rational" person (in the economic sense) has a utility function and a subjective probability measure, and makes decisions that optimize expected utility. But Arrow's impossibility theorem (look it up) states that (under any reasonable voting scheme) a democratic *group* may prefer A to B and B to C and C to A, which would imply $U(A) > U(B) > U(C) > U(A)$ (contradiction) if the *group* had a utility function. Political parties, companies, and entire countries can all be "irrational" to a greater extent than their individual members.

18.600 Probability and Random Variables

Fall 2019

Poisson

18.600 Problem Set 5

Welcome to your fifth 18.600 problem set! We'll be thinking more about Poisson random variables and the corresponding processes. Food for thought: should we replace the expression "Bad things happen in threes" with "When one bad thing is expected, the chance of exactly three is $1/6e$"?

A. FROM TEXTBOOK CHAPTER FOUR:

1. Theoretical Exercise 25: Suppose that the number of events that occur in a specified time is a Poisson random variable with parameter λ. If each event is "counted" with probability p, independently of every other event, show that the number of events that are counted is a Poisson random variable with parameter λp. Also, give an intuitive argument as to why this should be so. As an application of the preceding result, suppose that the number of distinct uranium deposits in a given area is a Poisson random variable with parameter $\lambda = 10$. If, in a fixed period of time, each deposit is discovered independently with probability $\frac{1}{50}$, find the probability that (a) exactly 1, (b) at least 1, and (c) at most 1 deposit is discovered during that time.

B. ANSWER THE FOLLOWING:

1. Let X be uniform on $[0, 1]$ and compute the expectation and variance of X^n where n is a positive integer.

2. Let X be uniform on $[0, 1]$ and compute the probability density function of $Y = X^3$.

C. In Regular Bus City, there is a shuttle bus that goes between Stop A and Stop B, with no stops in between. The bus is perfectly punctual and arrives at Stop A at precise five minute intervals (6:00, 6:05, 6:10, 6:15, etc.) day and night, at which point it immediately picks up all passengers waiting. Citizens of Regular Bus City arrive at Stop A at Poisson random times, with an average of 5 passengers arriving every minute, and board the next bus that arrives.

(a) Suppose that you visit this city and that you arrive at Stop A at a time chosen uniformly at random from the times in a day. How long do you expect to have to wait until the next bus?

(b) How many citizens of Regular Bus City do you expect to be on the bus that you take?

In Poisson Bus City, there is a shuttle bus that goes between Stop A and Stop B, with no stops in between. The times at which the bus arrives at Stop A are a Poisson point process with one bus arriving every five minutes on average, day and night, at which point it immediately picks up all passengers waiting. Citizens of Poisson Bus City (like those of Regular Bus City) arrive at Stop A at Poisson random times, with an average of 5 passengers arriving every minute, and board the next bus that arrives.

(c) Suppose that you visit this city and that you arrive at Stop A at a time chosen uniformly at random from the times in a day. How long do you expect to have to wait until the next bus?

(d) How many citizens of Poisson Bus City do you expect to be on the bus that you take?

(e) Are the following two statements true or false? If they are both true, explain in words the apparent discrepancy:

(i) When you visit, buses in Poisson Bus City seem on average to come twice as slowly and to be twice as crowded as those in Regular Bus City

(ii) In both cities, buses come on average every five minutes and people come on average five times per minutes, so that over the long haul there are 25 people per bus on average—so buses are on average equally crowded in the two cities.

Remark: Poisson Bus City is not the worst case scenario. Suppose that buses come in pairs (one right behind the other) with the pairs arriving as a Poisson point process with one pair every 10 minutes on average. And suppose that whenever this happens, everybody gets in the first bus and leaves the second bus empty. Now if you arrive at a random time, you can expect your bus to take four times as long to come and be four times as crowded as in Regular Bus City (assuming that like others you get on the first bus in a pair). On a real life bus route with many stops, the closer a bus is to the bus ahead of it, the faster it can go (since it is picking up fewer passengers) which can lead to this kind of clumping.

D. Each day (independently of all other days) Jill has a 1/2500 chance of hearing a particular fact: let's say the fact that Henry Mancini composed "The Pink Panther Theme." Jill stores something in long term memory after hearing it 3 times. Use Poisson approximations to (approximately) answer the following:

(a) What is the probability that, by the time Jill is 10,000 days old, she knows that Henry Mancini wrote "The Pink Panther Theme"?

Alice reads more than Jill and has a better memory for trivia. Each day (independently of all others) Alice has a 1/1000 chance of learning that Henry Mancini wrote "The Pink Panther Theme," and she stores information in long term memory after hearing it twice.

(b) What is the probability that, by the time Alice is 10,000 days old, she knows that Henry Mancini wrote "The Pink Panther Theme"?

(c) If there are $10,000$ similar facts (each fact comes with same probabilities as above), how many of them do we expect that Jill knows but Alice doesn't (assuming that both are 10,000 days old)? Assume that for each given fact, the two Poisson random variables (number of times fact is heard by Alice and by Jill) are independent. (If the answer is small, then Jill should feel pretty lucky when one of these facts comes up while she is watching Jeopardy with Alice.)

E. This problem addresses the Gompertz model for the duration of human life. But it starts out as another story about buses. Let X_1, X_2, X_3 be a Poisson point process of parameter 1 on $[0, \infty)$. Recall that this implies that X_1 and $X_2 - X_1$ and $X_3 - X_2$, etc., are i.i.d. exponential random variables each with parameter 1. Now for each integer $i \geq 1$, let $Y_i = \log X_i$. In Poisson

Bus City, you might imagine that a bus line starts operating at time zero, and thereafter bus arrivals correspond to the times X_i.

On Accelerating Frequency Planet (AFP) the bus arrival times are $Y_1, Y_2, \ldots$. That is, each arrivial time is the *natural logarithm* of a point in the Poisson point process. Time is measured from $-\infty$ to ∞ on AFP, so it is possible that some bus arrival times are negative.

(a) Show that, on AFP, given any constants $a < b$, the number of buses that arrive between times a and b is a Poisson random variable with parameter $\int_a^b e^x dx$.

(b) Explain (with a sentence each) why the following things are true: the number of buses that arrive during the time interval $(-\infty, 0]$ is Poisson with parameter 1, while with probability one *infinitely* many buses arrive after time 0. If $\alpha = \ln 2 \approx .7$, then for each k the expected number of buses that arrive during $[k\alpha, (k+1)\alpha]$ is twice as large as the the expected number that arrive during $[(k-1)\alpha, k\alpha]$. (In other words, the doubling time for the bus arrival-frequency rate is α.) Moreover, the *first* bus's arrival time is a random variable whose median is $\ln\big(\ln(2)\big) \approx -.37$.

Note you can approximate an ordinary Poisson point process with parameter λ by partitioning time into disjoint intervals of the form $[t, t+\epsilon)$ for small ϵ and asserting that each interval independently contains a bus with probability $\lambda\epsilon$. Things are similar on AFP, except that the probability is approximately $e^t\epsilon$ when ϵ is small; in some sense, this is like saying that the Poisson parameter λ is "time dependent" (and exponentially increasing) with $\lambda(t) = e^t$.

More to the story: at time -7 on AFP, an adorable but immobile sloth is born at the bus stop, where it lives until it is killed by the first bus that arrives. Since it is unlikely the first bus comes before time -7, (b) suggests that the sloth's life span is a random variable with median about $7 - .37 \approx 6.63$. The standard unit of time on AFP is the duodecennium (i.e., twelve years), so that α "units" means $12\alpha \approx 8.32$ years and the sloth's median life span is $6.63 * 12 \approx 80$ years.

(c) Suppose that on AFP, half of the buses have fat tires and half have thin tires (bus type decided by independent coin toss for each bus), and female sloths are only killed by fat tired buses, while males are killed by all buses. Argue that if the sloth is female, its life expectancy is about 8.32 years (i.e., α duodecennia) longer than if it had been male. (Hint: use problem A.1 and argue that the probability density function for the lifespan of a female born at time -7 agrees with that of a male born at time $-7 - \alpha$. Then note that having a bus between time $-7 - \alpha$ and -7 is very unlikely.)

(d) Let p_k be the probability that the sloth dies during its kth year of life, *given* that it has survived for $(k-1)$ years. Argue that p_k is approximately the expected number of buses that arrive during that kth year when p_k is small — say, less than .1. (This is related to arguing that the probability that a λ Poisson random variable equals 1 is approximately λ when λ is reasonably small.) Thus p_k grows (roughly) exponentially in k for the first 80 or so years of life.

(e) Look up `https://www.ssa.gov/oact/STATS/table4c6.html` and `https://en.wikipedia.org/wiki/Gompertz%E2%80%93Makeham_law_of_mortality` and (after looking them over for five minutes) write a sentence or two about what you noticed — and in particular about how closely the p_k corresponding to humans match those of the sloths on AFP. (Three obvious differences: humans are much more likely than AFP sloths to die during the first year or so of life. Humans in late teens and twenties — especially males — die at a rate that is higher than the Gompertz law would predict, hence the "bump" in the otherwise nearly straight line on the Wikipedia chart. This may be in part be due to risky behaviors, which are either more pronounced among people that age or simply make a larger difference on the log scale because other causes of death are low. Finally, outside the "bump" period the gap betwen death rates for male and female humans is large but not as large as for AFP sloths.) You may find it helpful to consult `https://gravityandlevity.wordpress.com/2009/07/08/your-body-wasnt-built-to-last-a-lesson-from-human-mortality-rates/` for a somewhat breezier account of the Gompertz law story. (Notice also that this problem has another part after the next few remarks.)

Remark: Gompertz law (i.e., exponential mortality rate growth) appears to apply pretty well to both humans and animals (with a species dependent doubling rate α). Google Gompertz morality and find out more. If you wanted a story to explain the exponential growth, one naive one would be that "glitches" in the body accumulate exponentially, and your death rate is proportional to the number of glitches. Another simplistic story is that if there is a genetic mutation that *causes* an organism to die at age X, then the rate at which natural selection eliminates that gene decreases with the proportion of organisms who reach age X. So mutations that cause functioning to break down in old people accumulate faster than those that affect young people. Note also that although Gompertz law holds pretty well in developed nations, it does not hold in settings where a large fraction of deaths are caused by predators, wars, infectious diseases, etc. that are as deadly to the young as the old.

Remark: The assumption of an exponential increase in "bus arrival frequency" suggests that an unhealthy habit that *doubles* your probability of dying within any given year should *subtract* about α units of life expectancy (one doubling period). Medical advances that eliminate half the number of fatal buses should *add* about α units of life expectancy. If medicine eliminated 7/8 of the buses, this should increase life expectancy by 3α units, about 25 years. Lifestyle choices that decrease the death rate by 29 percent (i.e., which multiply death rate during any given year by $1/\sqrt{2}$) should add about $\alpha/2$ units (about four years) to life expectancy (since your death rate each year is what it would be if you were 4 years younger).

Remark: *Life expectancy* is computed by estimating what fraction of people of each numerical age die each year, and then calculating how long newborns would expect to live if they died with the corresponding probablity each year. Per discussion above, doubling the death rate each year should *roughly* correspond to subtracting eight years of life expectancy. Decreasing the death rate by 29 percent *roughly* corresponds to adding four years of life expectancy. US life expectancy per `https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy` is 79.3 years.

Countries with life expectancy 8 years *lower* include Guatamela, Bangladesh, and Ukraine. Countries with life expectancy 4 years *higher* include Japan, Switzerland, and Singapore. Factors presumably include opioids, guns, cigarettes, obesity, alcoholism, pollution, diet, disease, etc.

Remark: In a world where Gompertz applies precisely with a doubling period of 8 years (and people's life spans are independent of each other), if you have $8 = 2^3$ students of age 20 and one professor of age $44 = 20 + 3 \times 8$, then the time until the first student dies should agree in law with the time until the professor dies. Essentially, being three doubling periods older means you expect eight times as many buses coming your way. But being eight people instead of one means that (as a group) you also have eight times as many buses coming your way. (If you want to extend the bus metaphor, imagine a different independent lane of buses for each person, with frequency rates depending on that person's age...) If you have a class of 128 students of age 20 and one older professor of age 76, then the time until the first student dies agrees in law with the time until the professor dies. If two parents are 24 years older than a child then, of the buses coming towards these three people, 1/17 are coming toward the child and 8/17 are coming toward each parent (since each parent has 8 times as many buses coming its way as the child). In fact, one can show that *given* the arrival times of the buses, we can assign each bus independently to one of the three people, with probabilities 8/17 for each parent and 1/17 for child. Looking at the first bus, we find that the probability that the child is the first of the three people to die is 1/17.

(f) Suppose that, in the world of the previous remark, there are four siblings of ages 20, 28, 28, and 36. What is the probability the oldest one dies first?

Remark: Per CDC, smoking roughly triples mortality `https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/tobacco_related_mortality/index.htm` and reduces life expectancy by over a decade. If this were true exactly (on our purely Gompertz 8-year-doubling world) then a smoker's death rate would be the same as that of a non-smoker who was $8 \cdot \log_2(3) \approx 12.67$ years older. If there were two young people of the same age (one smoker, one not) the non-smoker could say "You should not smoke. You are shortening your life by 12 years," and the smoker could counter "Yes, but don't be too sanguine about your own life span. There is still a 1/4 chance that I will outlive you." The point is that for both individuals the uncertainty is quite high. Thinking back to our original model, if X is a rate one exponential and $Y = 12(7 + \log X)$, then $\mathrm{SD}(Y) = 12\mathrm{SD}(\log X)$ which comes out to about 15 on wolframalpha. This is consistent with `https://www.nber.org/papers/w14093`, which states that the standard deviation of adult lifespan in the US is about 15 years. As the linked paper notes, this uncertainty makes it hard to plan for things like retirement and inheritance. C'est la vie.

Remark: To *drastically* increase human life span (so we live to 150, say) we'll have to change α — i.e., to slow the exponential growth of the mortality rate. How can we do this? Can we freeze blood and tissue from when we're younger and reintroduce it later? Can we reset age markers? Can we manually edit the genes that cause aging? Somebody in this class should figure this out. We have a lot of buses coming our way. (Google *naked mole rat Gompertz* for inspiration.)

18.600 Probability and Random Variables

Fall 2019

Exponentials and normal approximations

18.600 Problem Set 6

Welcome to your sixth 18.600 problem set! Let's warm up by thinking about data analyis. Imagine you are teaching a high school class and you give 30 students a multiple choice exam with 40 problems, and you come back with the following list of (rounded-down) percentage scores:

77, 75, 87, 82, 75, 85, 77, 62, 70, 85, 80, 82, 80, 72, 90,
90, 87, 80, 82, 72, 85, 90, 82, 75, 77, 75, 85, 65, 70, 85

Your educational data analyst might come back with the following observations about your class.

1. The three students with scores of 90 are unusually advanced.
2. The two students with scores of 62 and 65 are struggling.
3. The two students with scores of 70 are having at least some trouble and should be watched.
4. The two students with scores of 87 are among the stronger students and should be encouraged.
5. The above-80's are stronger than the below-80's. Dividing the class into two tracks might help.

This is the information you would convey at parent-teacher conferences, or to a guidance counselor who asked about class performance. But as it happens, in this *particular* example, the above assumptions are all false. The above numbers were created by a computer simulation, in which all students were *equally* capable, and each student solves each problem correctly and independently with probability 80 percent. We know that $\sqrt{npq} = \sqrt{40 \cdot .8 \cdot .2} = \sqrt{6.4} \approx 2.52$, and 2.52 problems corresponds to about 6.3 percentage points, so by de Moivre-Laplace we'd guess that a .68 fraction of students are between 73.7 and 86.3, which is *roughly* what we see. On the other hand, you can also imagine that the numbers correspond to something objectively measurable (like height, say) and that the students really are as different as the numbers indicate. It is hard to tell from the numbers alone. Nate Silver has a fun book about the challenge of distinguishing "signal from noise" in the real world. `https://www.amazon.com/Signal-Noise-Many-Predictions-Fail-but-ebook/dp/B007V65R54`

This problem set features problems about normal and exponential random variables, along with stories about coins, politics, and a fanciful bacterial growth model. We have not yet proved the central limit theorem, but we have presented a special case: the so called de Moivre-Laplace limit theorem, which already begins to illustrate why the normal distribution is so special.

A. FROM TEXTBOOK CHAPTER FIVE:

1. Theoretical Exercise 30: Let X have probability density f_X. Find the probability density function of the random variable Y defined by $Y = aX + b$.

REMARK: If you internalize the idea of the last problem (you understand how f_X is stretched, squashed, and translated when you replace X by $aX + b$) it makes it easier to understand and remember some of the formulas on the story sheet. Take a few minutes to stare at your answer and make sure it makes intuitive sense. You should appreciate why "multiplying X by a" corresponds to stretching graph of f_X horizontally by factor of a and vertically by factor of a^{-1}. And why adding b corresponds to translating graph by b unit to the right. Definitely make sure you understand what this means in the context of the first four continuous random variables on the story sheet.

B. At time zero, a single bacterium in a dish divides into two bacteria. This species of bacteria has the following property: after a bacterium B divides into two new bacteria B_1 and B_2, the subsequent length of time until B_1 (resp., B_2) divides is an exponential random variable of rate $\lambda = 1$, independently of everything else happening in the dish.

1. Compute the expectation of the time T_n at which the number of bacteria reaches n.
2. Compute the variance of T_n.
3. Are both of the answers above unbounded, as functions of n? Give a rough numerical estimate of the values when $n = 10^{30}$.

Remark: It may seem surprising that the variance is as small as it is. This is similar to radioactive decay models, where one starts with a large number n of particles, and the time it takes for the first $n/2$ to decay has a very small variance and an expectation that doesn't much depend on n — so that in chemistry we often talk about "half-life" as if it were a fixed deterministic quantity of time. In the example above, one can show that the variance of $T_{2n} - T_n$ is small when n is large (and that the expectation tends to a limit as $n \to \infty$) so we could talk about "doubling time" the same way.

C. In 2007, Diaconis, Holmes, and Montgomery published a paper (look it up) arguing that when you toss a coin in the air and catch it in your hand, the probability that it lands facing the same way as it was facing when it started should be (due to precession effects) roughly .508 (instead of exactly .5). Look up "40,000 coin tosses yield ambiguous evidence for dynamical bias" to see the work of two Berkeley undergraduates who tried to test this prediction empirically. In their experiment $20,245$ (about a .506 fraction) of the coins landed facing the same way they were facing before being tossed. A few relevant questions:

1. Suppose you toss $40,000$ coins that are truly fair (probably .5) and independent. What is the standard deviation of the number of heads you see? What is the probability (using the normal approximation) that the fraction of heads you see is greater than .506?

If X is the number of heads in a single fair coin toss (so X is 0 or 1) then X has expectation .5 and standard deviation .5. If $\tilde{X}$ is the same but with probability .508 of being 1 then $E[\tilde{X}] - E[X] = .008$. The quantity .008 is about .016 times the standard deviation of X (which is very close to the standard deviation of $\tilde{X}$). Suppose $Y = \sum_{i=1}^{N} X_i$, where the X_i are independent with the same law as X. Similarly suppose $\tilde{Y} = \sum_{i=1}^{N} \tilde{X}_i$, where the $\tilde{X}_i$ are independent with the same law as $\tilde{X}$.

2. Show that $E[\tilde{Y}] - E[Y]$ is $.016\sqrt{N}$ times the standard deviation for Y (which is approximately the same as the standard deviation of $\tilde{Y}$).

Note that if $N = 40,000$, we have $.016\sqrt{N} = 3.2$. So Y and $\tilde{Y}$ are both approximately normally distributed (by de Moivre-Laplace) with similar standard deviations, but with expectations about 3.2 standard deviations apart. The value the students observed is closer to the mean of $\tilde{Y}$ than to the mean of Y but the evidence for bias is not overwhelming.

3. Imagine that we had $N = 10^6$ instead of $N = 40,000$. How many standard deviations apart would the means of Y and $\tilde{Y}$ be then? Could you confidentally distinguish between an instance of Y and an instance of $\tilde{Y}$?

Remark: In this story, X and $\tilde{X}$ have about the same standard deviation and $d = (E[\tilde{X}] - E[X])/\mathrm{SD}[X] = .016$. This ratio is sometimes called *Cohen's d*. (Look this up for a more

precise definition.) This ratio is a good indication of how many trials we would need to *detect* an effect. If you did N trials and you had $\sqrt{N}d > 10$ then you could detect the effect very convincingly with very high probability. In practice it is often hard to do $N = 100/d^2$ independent trials when d is small. Moreover, even if we found the research budget to toss $400,000$ coins, we would not know whether coins tossed in real life scenarios (e.g. sporting events) had the same probabilities as coins tossed by weary researchers doing hundreds in a row.

Remark: The third significant digit of a coin toss probability may seem unimportant (albeit undeniably interesting). But imagine that every year 10^6 people worldwide have a specific kind of heart attack. There is one treatment that allows them to survive with probability .5 and another that allows them to survive with probability .508. If you could demonstrate this and get people to switch to the second treatment, you could save (in expectation) thousands of lives per year. But as a practical matter it might be impossible to do a large enough controlled trial to demonstrate the effect. It is (to put it mildly) harder to arrange a randomized experiment on a heart attack victim than it is to toss a coin.

Remark: You might even have trouble distinguishing between a treatment that gives a .4 chance of survival and one that gives a .6 chance. Yes, a trial with a few thousand people would overwhelmingly demonstrate the effect (and a trial with 100 people would *probably* at least *suggest* the right answer) but there is no guarantee that the right kind of clinical trial has been (or even can be) done — or that your busy doctor is up to date on the latest research (especially if your condition arises infrequently). Collecting and utilizing data effectively is a huge challenge.

D. In Open Primary Land, there are two political parties competing to elect a senator. There is first a *primary election* for each party to select a nominee. Then there is a *general election* between the two party nominees. A voter can vote in either party's primary, but not in both. Suppose that A_1 and A_2 are the only two viable candidates in the first party's primary and B_1 and B_2 are the only two viable candidates in the second party's primary. Let $P_{i,j}$ be the probability that A_i would beat B_j if those two faced each other in the general election. Let $V(A_1), V(A_2), V(B_1), V(B_2)$ be the *values* you assign to the various candidates, and assume that your sole goal is to maximize $E[V(W)]$ where W is the overall election winner.

1. Check that $V(A_i, B_j) := P_{i,j}V(A_i) + (1 - P_{i,j})V(B_j)$ is the expectation of $V(W)$ *given* that A_i and B_j win the primaries.

Now, to determine your optimal primary vote, you need only figure out how to maximize $E[V(A, B)]$, where A and B are the primary winners. Assume that (aside from you) an even number of people vote in each primary (with fair coin tosses used to break ties).

2. Argue that if you vote for candidate A_1 the expected value of your vote is

$$\frac{1}{2}p_1\big(V(A_1, B_1) - V(A_2, B_1)\big) + \frac{1}{2}p_2\big(V(A_1, B_2) - V(A_2, B_2)\big)$$

   where p_i is the probability that B_i wins the second primary *and* the first primary voters are tied without you, so that your vote swings the election to A_1. (To explain the $\frac{1}{2}$ factor, recall that a coin toss takes your place if you don't vote.) You can compute values for other candidates similarly. You want to maximize your vote's expected value.

3. Argue that the expected value of voting for A_2 is minus one times the expected value of voting for A_1 (similarly for B_1 and B_2).

4. Argue that if you replaced V with $-V$ then your choice of *which primary* to vote in would stay the same, but your choice of *which candidate* to vote for would change.

Remark: The result of (d) suggests that a far-right voter (who just wants to pull the country as far right as possible) and a far-left voter (who just wants to pull the country as far left as possible) should actually vote in the *same* primary. Roughly speaking, they find the primary in which a vote makes the most marginal difference and they both vote there (albeit for different candidates). This may seem surprising, because many people assume that far-right voters should always vote in the further right party's primary and that far-left voters should always vote in the further left party's primary (even when rules explicitly encourage voters to vote in whichever primary they like). There are no doubt be many reasons for this, but part of the reason may be that calculating the expected impact of a primary vote is *complicated* and *unintuitive*. Perhaps somebody should make an app so that you just plug in $V(A_1), V(A_2), V(B_1), V(B_2)$ (perhaps normalized so that your favorite candidate has score 100 and your least favorite has score 0) and the app estimates the relevant probabilities from prediction markets and polls and tells you how to vote. In the meantime, the simple "vote for the candidate you like most" strategy seems likely to remain popular.

Remark on *reasons* for things: If you toss 101 fair coins, a binomial calculation shows that there is about a .15 chance that the number of heads will be 50 or 51, so that a heads vs. tails majority vote *comes down to one vote*. If, for example, there turn out to be exactly 50 heads, you can say that *any* of the 51 tails votes *could* have swung the election outcome if had they voted differently. So it may be technically accurate, albeit misleading, to say "Heads lost because the 7th coin was tails" *and* "heads lost because the 19th coin wasn't heads" *and* "tails won because the 78th coin was tails" and so forth. If you google the phrases "won because" and "lost because" (or "didn't win because" and "didn't lose because") in quotes you'll find lots of similarly dubious attempts to declare that certain factors in close political elections and sporting events were or weren't *the reason*. Of course, when a contest is close, it may be accurate (if banal) to say nearly every factor was decisive. Yet humans seem oddly attached to the idea that things happen for *specific* reasons. (Any specific reason for this?)

E. Harper and Heloise are real estate agents for a corporate firm. Once a week, each of them is assigned to close an important deal. It is known that one of the two associates closes her deals successfully 60 percent of the time (model these as i.i.d. coin tosses) and the other 50 percent (also i.i.d. coin tosses) but you are not sure which is which. You formulate a plan: you will wait N weeks, so that each associate gets to attempt N different deals, and then you will offer a permanent job to the associate who is ahead in number of closings. The **main question** we'd like to answer is this: roughly how large does N have to be to ensure that there is a 95 percent chance that the more capable closer (i.e., the one with closing probability .6) is ahead after N steps? We'll approximately solve this in three steps:

1. Let X_N and Y_N be the number of deals closed by (respectively) the more and less capable agents agent after N steps. So X_N and Y_N represent the number of heads in N tosses of a p-coin with (respectively) $p = .6$ and $p = .5$. Compute (in terms of N) the mean and variance of the random variable $S_N = X_N - Y_N$.

2. For the random variable S_N, compute (in terms of N) how many standard deviations 0 is below the mean. That is, find $E[S_N]/SD[S_N]$ where SD denotes standard deviation.

3. The De Moivre Laplace theorem (special case of the central limit theorem, which will come later in the course) suggests that if N is large, both X_N and Y_N are approximately normal variables.

Since X_N and Y_N are independent (and since the difference between two independent normal random variables is itself normal) one can argue that $S_N = X - Y$ is also roughly Gaussian. (You don't have to formally prove this. Just take it as given for now.) In particular, if Z_N is a normal random variable with the same mean and variance as S_N then $P(S_N > 0) \approx P(Z_N > 0)$. Compute an approximate value for $P(Z_N > 0)$ when $N = 143$. We can interpret this as an approximation for the probability that S_N is positive (so the better closer wins). If it helps, you may assume that $P(X \leq 1.7) \approx .95$ for normal X with mean zero and variance one and that $\sqrt{143}/7 \approx 1.7$. Conclude that 143 is roughly the answer to the main question.

Remark: Even though there is a *huge* difference between the two agents, it actually takes *years* to determine with confidence which is better. If you as the manager *think* you can tell based on just a few outcomes, you are deluding yourself—the noise to signal ratio is too high. This problem appeared (without the real estate agent story) in the 538 Riddler `http://fivethirtyeight.com/features/rock-paper-scissors-double-scissors/` (which often has great probability puzzles) and also in an academic paper `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3034686` which surveyed financial experts to see how many flips they thought were necessary. Feel free to look up these references for more detailed calculations. The paper states:

"The median guess was 40 flips. While lower than the full-credit answer of 143, it does show that the respondents in general appreciate it takes a long time to identify a phenomenon with this kind of risk/reward ratio simply by history. We include in Appendix 1 the calculation used to arrive at 143.3 Our respondents are a pretty mathematical bunch, and we suspect that if they took their time to calculate an answer, rather than giving a quick guess as we requested, most would have arrived at the correct answer. But the point of the exercise was to illustrate how when we are thinking fast, we tend to overweight the value of small samples: a full 30% of respondents, the single largest bucket, thought 10 flips or less was sufficient. This built-in bias to over-weight small samples results in a tendency to ignore the investing dictum 'past performance is not indicative of future results' when we clearly should not."

I am not sure whether real estate agents employ this particular strategy when deciding who to hire. But marketers of all kinds regularly do something called "A/B testing" or "split testing" where they run two versions of an ad for a period, and then settle on the one that leads to the most clicks (or the most "conversions," whatever that means in the context — purchases, subscriptions, likes, etc.) You could argue that this is one of the very simplest kinds of machine learning. Google *A/B testing* to read more.

Remark: Economics Planet has two political parties. When one is in power, the economy is good with probability .5. When the other is in power, the economy is good with probability .6. The second party is then much better for the economy on average, but it would take over a thousand years (of alternating parties every 4 years) to be 95 percent confident that we could determine which was which.

Remark: It is fun to think of other stories along these lines. Maybe two students get only A or B grades, but one has A probability .5 and the other .6. Can you tell which is which based on GPA? Or maybe one medicine cures your headache with probability .5, and the other with probability .6. Or one airline has good food with probability .5 and the other with probability .6. Or one journal accepts your academic papers with probabilty .5 and one with probability .6. Each such story is a parable about the difficulty of learning from experience in the absence of large data sets.

Remark on preconceptions: Let H be the event that Harper is the stronger candidate and T the

event that Harper closes more deals during the first 143 trials. Suppose that we think *a priori* (based on resumes, interviews, the fact that Harper went to MIT, etc.) that $P(H) = .95$. Since we know (approximately) that $P(T|H) = .95$ and $P(T|H^c) = .05$ we can deduce (using the Bayesian analysis we did for disease trials) that $P(H|T) = .5$. That is, even after learning that Harper was behind after three years of data, we still think there is a .5 chance that Harper is stronger. Similarly the political partisans of Economics Planet, who start out thinking one party is highly likely to be better for the economy, may not fully reverse their opinions even after they learn that the opposing party did better over a 1000 year period.

Remark on smaller samples: We need $N = 143$ tosses for 95 percent confidence, but we still learn *something* when $N < 143$. Suppose $N = 1$ for Harper and Heloise: so if exactly one person closes a deal the first week, we give the job to that person; otherwise we toss a fair coin to see who gets the job. In this case, one can show that the stronger candidate gets the job with probability .55 (which is better than the .5 we'd have if we just guessed without considering first week performance). With a year of data (52 tosses), the stronger candidate wins with over 80 percent probability.

Remark on baseball: A baseball player might have over 500 at bats during a season. So (based on results from this problem) it is possible to distinguish between a .400 hitter and a .500 with 95 percent probability after less than a third of a season. But with one season worth of data, you cannot distinguish (with 95 percent probability) between a .253 hitter and a .286 hitter. These are the batting averages corresponding to 25th and 75th percentile players according to `https://www.fangraphs.com/library/statistic-percentile-charts`. Does this disturb any baseball fans in this course?

F. Imagine a simplified game of basketball in which the game is played until exactly 101 shots are made (each worth 2 points) and the winner is the team that has made the most shots. Let X_n be 1 if the nth shot to be made is made by the first team, and 0 otherwise. Assume $X_1, X_2, \ldots, X_{101}$ are independent, each equal to 1 with probability .52 and 0 with probability .48. (So the first team is a bit stronger.) Use a calculator like `https://stattrek.com/online-calculator/binomial.aspx` or `wolframalpha.com` to give numerical appproximations for the following:

1. The probability that the first team wins the game.
2. The probability that the first team wins at least four games out of a series of seven independent games like the one described above.
3. The probability that the first team makes more shots *total* than the second team over the course of the seven games (i.e., makes at least 354 of the 707 total shots). Is this higher or lower than the number from the previous answer? Give an intuitive explanation for the difference.

Remark: At this point in the problem set, are you surprised that such a small point-by-point advantage translates into such a large advantage overall? Or are you surprised the other direction, i.e., surprised that even with 707 points played, the stronger team has a non-negligible chance of losing? On another note, why do you think so many sports decide the winner of a series by counting the number of games won instead of the cumulative number of points? Is it because it gives underdogs more of a chance? Or does it make the games more fun to watch? Note that in actual basketball, the X_i are not independent — since when one team makes a shot, the other team gets possession and is more likely to make the next shot. We could model possession with Markov chains (coming later in the course). Or we can imagine that in our simplified game, possession after each shot is decided by a "jump ball," so that the independence assumption is more plausible.

18.600 Probability and Random Variables

Fall 2019

Cauchy, beta, and gamma random variables
18.600 Problem Set 7

Welcome to your seventh 18.600 problem set! This problem set features problems about beta, Gamma, and Cauchy random variables. These random variables are not quite as ubiquitous as others we have discussed (exponential, uniform, normal, Poisson, binomial) but they are fun and do come up. The problems should help you internalize the definitions and some of the standard interpretations. In particular we will have several problems exploring the idea of beta distributions as Bayesian posteriors for the p associated to a biased coin. Doing these problems will also give you a chance to think a bit more about things we have done earlier in the course (expectation, joint distributions, conditional probability, etc.) The idea is that you initially think that p is a uniform random variable in $[0, 1]$. But then you see the outcomes of a few coin tosses (e.g., maybe you see "heads, heads, tails, heads, heads") and you revise your opinion about what p is likely to be; and *given* that you have seen $(a-1)$ heads and $(b-1)$ tails, you now think that p is a beta random variable with parameters a and b. (You might need to review the lecture notes and/or textbook discussion on beta random variables.)

A. FROM TEXTBOOK CHAPTER FIVE: Theoretical Exercise 26: If X is a beta random variable with parameters a and b show that

$$E[X] = \frac{a}{a+b},$$

$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

B. Suppose p is a random variable that takes values in $[0, 1]$ and has a density function f defined on $[0, 1]$. Imagine a two part experiment where we first choose p from this distribution and *then* we toss a p-coin k times. Let $X_i \in \{T, H\}$ be the value of the ith toss.

1. Here is an alternative setup. Take p as above and let Y_i be uniform random variables on $[0, 1]$, where the Y_i and p are independent. Set $X_i = \begin{cases} H & Y_i \leq p \\ T & Y_i > p \end{cases}$. Explain why the p and X_i defined this way have the same joint probabilistic law as the p and X_i defined above.
2. Write down the joint density function of p and Y_1 on $[0, 1]^2$ and use it to argue that $P\big(X_1 = H\big) = E[p]$.
3. Compute $P\big(p \leq a \; X_1 = H\big)$. Compute the derivative of this quantity (as a function of a) and argue that what you get can be interpreted as the *posterior* probability density function for p *given* that $X_1 = H$. Show that this function is (some constant times) the function $x \to xf(x)$. In other words, seeing a heads causes you to revise your probability density for p by multiplying it by x — and then by a constant to make it so the total integral is still one. Explain in a sentence why this makes sense intuitively.

Remark: If we see T instead of H we multiply by $(1 - x)$ instead of x. This is where the beta distribution comes from: you start with the uniform distribution $f(x) = 1$ and multiply by x for each heads and $(1 - x)$ for each tails, always multiplying by constant to make the total integral one. The set $\{T, H\}^n \times [0, 1]$ of outcomes for $(X_1, X_2, \ldots, X_n, p)$ is a union of 2^n copies of $[0, 1]$. The ideas above give us a way to describe a probability density function on that union of intervals.

C. Let p be the fraction of MIT students who love Marvel movies — or, more precisely, the fraction who will *say* they love Marvel movies when you ask (making it clear that you *absolutely* require a simple yes or no answer). Let's make believe that your initial Bayesian prior for p is uniform on $[0, 1]$. Now ask three of your fellow students (actually do this!) one at a time whether they love Marvel movies, and write the pair (# yes answers so far, # no answers so far) before you start and after each time you ask a question. For example, you will write the pairs

$$(0, 0), (1, 0), (2, 0), (3, 0)$$

if everyone you ask loves Marvel movies. Pretend that you have chosen your people uniformly at random from the large MIT population, so that each answer is yes with probabilty p and no with probability $(1 - p)$ independently of the other answers. Then write down each of the four number pairs, and beside each one draw a rough picture of the graph of the revised probability density function for p that you would have at that point in time, along with its algebraic expression, which should be a polynomial whose integral from 0 to 1 is 1. You can use graphing software if you want. Beside each graph write down the corresponding conditional expectation for p (using the results from part A) given what you know at that time.

Remark: Previous years asked this question about Taylor Swift and Ariana Grande. If you have a recommendation for next year's question, let me know. :) The ideal is someone or something well known but whose popularity within MIT is something most of us would be *a priori* unsure of, so that the uniform prior doesn't seem *too* unreasonable. (I have seen enough opinion polls to suspect that very few national *politicians* are loved by more than 70 percent of a population...)

D. Let X be a gamma random variable with parameters $\lambda = 1$ and n equal to some positive integer. Compute the expectation $E[X^k]$ in two ways:

1. Recall that X has the same law as $X_1 + X_2 + \ldots + X_n$ where X_i are i.i.d. exponential random variables, each with parameter $\lambda = 1$. Hence $E[X^k] = E[(X_1 + X_2 + \ldots + X_n)^k]$. If we expand $(X_1 + X_2 + \ldots + X_n)^k$ we get n^k terms that each look, for example, something like this: $X_4X_3X_1X_4X_3X_1X_7$. More precisely, each term is an ordered product of k factors, and each of the k subscripts can be any number from 1 to n. Given *one* of those terms (i.e., one of these ordered products) let a_j be the number of times the jth subscript appears. So $\sum_{j=1}^{n} a_j = k$.
   (a) Compute how many possible a_j sequences there are using stars and bars.
   (b) Compute the number of terms corresponding to a fixed a_j sequence. (This should be some multinomial coefficient.) Call this number A. (It depends on the a_j sequence.)
   (c) Compute the expectation of a single term corresponding to that sequence (i.e., a single product with a_1 subscripts given by 1, a_2 subscripts given by 2, etc.) This should be some product of factorials. Call this number B and compute AB to get the expectaton of the sum of all terms that corresponding to the given a_j sequence.
   (d) Then sum the AB product (whatever it comes out to be) over all possible a_j sequences.
2. Just write down the density function f_X and compute $E[X^k] = \int_0^\infty f_X(x)x^k dx$ as an integral. This should be easier and should (if all goes well) give you the same answer.

E. Alice and Bob are interested in having a child and, after difficulty conceiving, decide to undergo a medical procedure called IVF. In their universe, each couple has a random quantity p, uniformly distributed on $[0, 1]$, which indicates the probability that they will conceive a child after a cycle of IVF treatment. (The value p depends on permanent biological characteristics of Alice and Bob, but its value is unknown to them, so we model it as a random variable.) If Alice and Bob attempt multiple cycles, each one succeeds with the same probability p, independently of what happens on previous cycles.

(a) Explain intuitively why (in this universe) the probability that Alice and Bob conceive after one cycle should be .5 (i.e., the expected value of p).

(b) *Given* that Alice and Bob did not conceive during the first $(k-1)$ cycles, what is the updated Bayesian probability density for the random variable p?

(c) Use the answer in (b) to explicitly compute the expected value for p, given that the couple did not conceive during the first $(k-1)$ cycles. The answer is the conditional probability that the couple conceives during the kth cycle, given that they did not conceive during the first $(k-1)$ cycles. (As noted earlier on this problem set, one can prove in general that if one *first* chooses r in some random fashion and *then* tosses a coin that is heads with probability r, the overall probability of heads is the expectation $E[r]$.)

(d) Compute the conditional probability describe in (c) in a different way: imagine that $X_0, X_1, X_2, \ldots, X_k$ are uniformly and independently distributed on $[0, 1]$. Write $p = X_0$ and declare that the jth cycle succeeds if and only if $X_j < X_0$. Show that this model is equivalent to the one initially described, and then explain why the probability that $X_k < X_0$, *given* that X_0 is the smallest of the set $\{X_0, X_1, \ldots, X_{k-1}\}$, should be $1/(k+1)$. [Hint: use symmetry to argue that *a priori* the rank ordering of $X_0, X_1, \ldots, X_k$ is equally likely to be given by each of the $(k+1)!$ possible permutations.]

(e) Suppose that instead of being uniform the random variable p is *a priori* distributed on $[0, 1]$ according to the density function $f(x) = 2 - 2x$. (This might be more realistic, see remark below.) Under this assumption, compute the probability of success on the kth cycle given that the first $(k-1)$ cycles failed. [Hint: recognize $f(x)$ as itself a beta random variable and reduce to the previous case.]

Remark: This problem was inspired by a NY Times article called *With in vitro fertilization persistence pays off* (look it up) which reports on a large study:

> The rate of live births for participants after the first cycle in the new study was 29.5 percent, compared with 20.5 percent ater the fourth cycle, 17.4 percent after the sixth cycle, and 15.7 percent after the ninth cycle.

The numbers start a bit below our answer in (e) (since $.295 < 1/3$) and end up larger (since $.157 > 1/11$). This may suggest that p values are not distributed according the f that we guessed (somewhat arbitrarily) in (e). On the other hand, maybe different people have different Bayesian priors for p (based on age, known physical issues, etc.) and those whose p values are expected a

priori to be small tend to discontinue IVF after fewer cycles; if so, this could explain the higher reported success rates for later cycles.

F. Suppose $X_1, X_2, \ldots, X_{10}$ are independent Cauchy random variables. Compute the probability that $\sum_{i=1}^{5} X_i < 10 + \sum_{j=6}^{10} X_i$. (Hint: try combining the spinning flashlight story with left-right symmetry and the fact that the average of independent Cauchy random variables is itself a Cauchy random variable.)

G. On Planet A a site called rottentomatoes.com analyzes movie reviews. Each review is classified "fresh" if it seems on balance positive, "rotten" otherwise. Each movie has an *a priori* quality parameter $p \in [0, 1]$. After it is released, professional reviewers arrive one at a time and write reviews, each of which is fresh with probability p (independently of the others). The Tomatometer Score is the overall percentage of reviews that were fresh, expressed as a number between 0 and 100. One can show — using the *strong law of large numbers*, which will appear later in this course — that no matter what p is, the Tomatometer Score will (with probability one) converge to $100p$ in the limit as the number of reviews tends to infinity; so e.g. if $p = 3/5$ then the Tomatometer score will converge to 60 in the long run.

1. Suppose one movie has quality parameter .5 and another .6. Use normal approximations to estimate the probability the former gets a higher Tomatometer score than the latter after each movie has 143 reviews. (Hint: remember Harper and Heloise.)

2. Repeat the above with one movie having parameter .8 and the other .9, and with 100 reviews for each movie. (In both this problem and the previous one, the higher quality movie *probably* scores higher, but in neither case is it a sure thing.)

3. Imagine a studio makes a movie but has no idea in advance how well it will be received. The movie has a quality parameter p, but the studio does not know what it is and *a priori* considers p to be a *uniformly random variable* on $[0, 1]$. But then reviewers arrive one at a time to make reviews, each rating the movie fresh with probability p and rotten otherwise. Using beta random variables, give the *conditional* probability density for p given that one has seen f fresh and r rotten reviews so far.

4. Argue that if the studio does not know p, and knows only the number of f and r reviews seen so far, then it considers the probability of the *next* review being fresh to be $\frac{(f+1)}{(f+1)+(r+1)}$. (You can use the things derived in the IVF problem.) Using this compute the probability that the first four reviews are fresh, rotten, fresh, fresh in that order.

On Planet B, each released movie is initially given one fresh and one rotten review (to get the ball rolling). After that reviewers arrive one at a time to write and post reviews. But these reviewers do not form opinions independently; instead, each reviewer selects, uniformly at random, one of the *previously posted* reviews and writes a review of the same type (fresh or rotten). Let F_n be the fraction of the first n reviewers who rated a movie fresh. (We know $F_2 = 1/2$, but F_3 could be 1/3 or 2/3, and F_4 could be 1/4, 2/4 or 3/4.) Ultimately an infinite number of reviewers arrives, and the Tomatometer score is the limit $\lim_{n\to\infty} 100F_n$.

5. What is the probability on this planet that the first four reviews (after the "get the ball rolling" two) are fresh, rotten, fresh, fresh in that order? Does your answer agree with the answer computed above for the same sequence on Planet A? Would this still be be true if we replaced "fresh, rotten, fresh, fresh" by *any* finite length sequence?

Remark: It seems oddly coincidental that each sequence has the same probability on Planet A as on Planet B, even though the mechanism for generating the sequence is *completely* different.

6. Use comparison to Planet A to argue that on Planet B the limiting Tomatometer score is a uniformly random variable on $[0, 100]$.

Remark: On Planet A, you can imagine that a sufficiently skilled movie expert could figure out p after seeing an advance screening of the movie. This expert would then know *exactly* what the Tomatometer score would converge to in the $n \to \infty$ limit. But on Planet B, it is impossible to know anything at all about the limiting score just from seeing the movie.

Remark: Are the mechanisms of *our* world is closer to A (where reviewers see same movie but otherwise work independently) or B (where reviewers influence each other, and final consensus is unrelated to quality)? What explains why *Mona Lisa* and *Starry Night* are such iconic art works and *Baby Shark* has nearly 4 billion views? I have no answer, but I include a story below.

Quoted Remark (from Cass R. Sunstein's book *On Rumors*): The Princeton sociologist Matthew Salganik and his coauthors 14 created an artificial music market among 14,341 participants who were visitors to a website that was popular among young people. The participants were given a list of previously unknown songs from unknown bands. They were asked to listen to selections of any of the songs that interested them, to decide which songs (if any) to download, and to assign a rating to the songs they chose. About half of the participants made their decisions based on the names of the bands and the songs and their own independent judgment about the quality of the music. This was the control group. The participants outside of the control group were randomly assigned to one of eight possible subgroups. Within these subgroups, participants could see how many times each song had been downloaded. Each of these subgroups evolved on its own; participants in any particular world could see only the downloads in their own subgroups.....

It turned out that people were dramatically influenced by the choices of their predecessors. In every one of the eight subgroups, people were far more likely to download songs that had been previously downloaded in significant numbers—and far less likely to download songs that had not been so popular. Most strikingly, the success of songs was highly unpredictable. The songs that did well or poorly in the control group, where people did not see other people's judgments, could perform very differently in the "social influence subgroups." In those worlds, most songs could become very popular or very unpopular, with everything depending on the choices of the first participants to download them.

7. Imagine that on Planet B we "get the ball rolling" using a positive reviews and b negative reviews (instead of one of each). Can you generalize the argument used in the previous question to show that the limiting score is (100 times) a beta random variable in this case? Are the parameters just a and b? Google *Pólya's urn* for more on this model.

18.600 Probability and Random Variables

Fall 2019

Correlation, regression, and paradox

18.600 Problem Set 8

Welcome to your eighth 18.600 problem set! Let's think about correlations. Given a population of people who each have two attributes, like ACT and SAT scores, you can choose a member at random and interpret the attributes of the person you choose as random variables. These variables have a correlation coefficient, which you'd expect to be high for ACT and SAT scores (maybe about .87, per some site I googled).

How high is a .87 correlation really? Imagine X, Y_1 and Y_2 are independent standard normals. Imagine that X is your "raw test-taking acumen" and $X + aY_1$ is your score on one test and $X + aY_2$ is your score on another test, where the aY_i are random noise terms. Then $\rho(X+aY_1, X+aY_2) = 1/(1+a^2)$, which is about .86 if $a = .4$. In this case, the standard deviation of the "noise factor" is .4 times the standard deviation of the "raw ability factor," which may seem like quite a bit of noise (despite the fact that .86 seems pretty close to 1). Play the game `http://guessthecorrelation.com/` to get a sense of what different correlation levels look like.

Correlations inform beliefs. The strong observed correlations between cigarettes and early death (and many specific ailments like lung cancer) are a huge part of the of evidence that cigarettes are unhealthy. The discovery of the unhealthiness of smoking has saved millions of lives — a win for observational statistics. (Also an embarrassment, given that it took until the second half of 20th century for the case to be made persuasively.)

On the other hand, we all know the *correlation does not imply causation* cliché. The "spurious correlation" website tylervigen.com illustrates this with strong correlations between seemingly unrelated annual statistics like sociology doctorates and space launches, or pool drownings and Nicolas Cage films. The 2012 NEJM article *Chocolate consumption, Cognitive function, and Nobel Laureates* (which presents a real country-to-country correlation between chocolate consumption and Nobel prize winning) is a clever parody of the way observed correlations are used in medicine (it's only three pages; look it up). It earnestly walks the reader through causal hypotheses (brain-boosting flavanoids), reverse causal hypotheses (chocolate consumed at Nobel prize celebrations) and common demoninator hypotheses (geography, climate, socioeconomics), mostly dismissing the latter.

Alongside *correlation does not imply causation* one might add *correlation does not imply correlation*, or more precisely, reported correlation in some data set does not imply correlation in the larger population, or correlation that will persist in time. Google *study "is linked to"* and scroll through a few pages of hits. Some sound fairly plausible ("walking/cycling to work linked to lower body fat") but others raise eyebrows. Clicking through, you find that sometimes the correlations are weak, the sample sizes small, the stories (at least at first glance) far fetched. A news organization's criteria for deciding which links to publicize may differ from those a careful scientist would use to decide what to take seriously and/or study further (e.g. with randomized trials). Reader beware.

This problem set will also feature moment generating functions, regression lines (which you will encounter often in life) and a phenomenon called regression to the mean.

On a rather different note, many of you are familiar with *Pascal's wager*. The idea is that if choosing A over B comes with a finite cost but a positive probability (however small) of an infinite payoff, then one should always choose A. Pascal's conclusion was that if living a virtuous life leads (with even a tiny probability) to an eternal reward, then it is a worthwhile sacrifice to make. A common criticism is that this kind of thinking can lead to violence (killing heretics who *might* lead souls astray, or dissidents who

might obstruct an endless Marxist utopia) as well as virtue. A more mathematical concern is that in principle there may be many choices, each of which we expect to do an infinite amount of good (and perhaps also an infinite amount of harm) and that there is no obvious mathematical way to compare the competing infinities.

The comparison difficulties associated with infinite expectations can arise even when the payoffs themselves are finite with probability one (e.g., if the utility payout is a Cauchy random variable). This problem set illustrates this point with a particularly vexing form of a famous envelope switching paradox. Interestingly, in this paradox, the conditional expectations used for decision making are all finite; but a certain *a priori* expectation is infinite, and that is the root of the paradox. I hope that you enjoy thinking about the story, and that it causes you at most a finite amount of existential angst.

A. FROM TEXTBOOK CHAPTER SEVEN:

1. Problem 51: The joint density of X and Y is given by $f(x,y) = \frac{e^{-y}}{y}$, $0 < x < y$, $0 < y < \infty$. Compute $E[X^3|Y = y]$.

Remark: The next problem will help solidify your understanding of moment generating functions. These play a central role in *large deviation theory*, which in turn plays a central role in information theory, data compression, and statistical physics. In this course, we mostly use moment generating functions (and the closely related *characteristic functions*) as tools for proving the central limit theorem and the weak law of large numbers.

2. Theoretical Exercise 48: If $Y = aX + b$, where a and b are constants, express the moment generating function of Y in terms of the moment generating function of X.

B. Suppose that X and Y both have mean zero and variance one, so that $E[X^2] = E[Y^2] = 1$ and $E[X] = E[Y] = 0$.

(a) Check that the correlation coefficient between X and Y is $\rho = E[XY]$.

(b) Let r be the value of the real number a for which $E[(Y - aX)^2]$ is minimal. Show that r depends only on ρ and determine r as a function of ρ.

(c) Check that whenever Z has finite variance and finite expectation, the real number b that minimizes the quantity $E[(Z - b)^2]$ is $b = E[Z]$.

(d) Conclude that the quantity $E[(Y - aX - b)^2]$ is minimized when $a = r$ and $b = 0$.

Remark: We have shown that among all affine functions of X (i.e. all sums of the form $aX + b$ for real a and b) the one that best "approximates" Y in (in terms of minimizing expected square difference) is rX. This function is commonly called the *least squares regression line* for approximating Y (which we call a "dependent variable") as a function of X (the "independent variable"). If $r = .1$, it may seem odd that $.1X$ is considered an approximation for Y (when Y is the dependent variable) while $.1Y$ is considered an approximation for X (when X is the dependent variable). The lines $y = .1x$ and $x = .1y$ are pretty different after all. But the lines are defined in different ways, and when

$|r|$ is small, the correlation is small, so that neither line is an especially *close* approximation. If $r = 1$ then $\rho = 1$ and both lines are the same (since X and Y are equal with probability one in this case).

Remark: The above analysis can be easily generalized (by simply rescaling and translating the (X, Y) plane) to the case that X and Y have non-zero mean and variances other than one. The subject known as *regression analysis* encompasses this generalization along with further generalizations involving multiple dependent variables, as well as settings where a larger collection of functions plays the role that affine functions played for us. Regressions are ubiquitous in academic disciplines that use data. Given data in a spreadsheet, you can compute and plot regression lines with the push of a button (find a chart online; copy or import it into a free spreadsheet like sheets.google.com; control click letter on top of two distinct columns to highlight them; click the chart icon; then Customize, Series, Trendline... or if that doesn't work google *spreadsheet regression* for better instructions; or type "linear fit $\{1, 3\}\{2, 4\}\{4, 5\}\{3, 5\}$" into wolframalpha to see a simple example). For a more difficult setting, imagine that I give you a set of pictures, and assign a number to each picture indicating how closely it resembles a cat. If you had a nice way to approximate this function (from the set of digitally encoded pictures to the set of numbers) you could program it into your computer and enable your computer to recognize cats. Your procedure for approximating this function will likely be more complicated than a simple regression — it may involve *neural nets* or other tools from *machine learning*. Statistics and machine learning are hot topics, and may be part of your further coursework at MIT. (Note: even if your cat recognizing algorithm is a sophisticated neural net refined by hundreds of tinkering MIT alumni, you will still use the math behind the "clinical trial" stories in this course when you try to *test* your algorithm's effectivness.)

C. On Smoker Planet, each person decides at age 18, according to a fair coin toss, whether or not to become a life long cigarette smoker. A person who does not become a smoker will never smoke at all and will die at a random age, the expectation of which is 75 years, with a standard deviation of 10 years. If a person becomes a smoker, that person will smoke exactly 20 cigarettes per day throughout life, and the expected age at death will be 65 years, with a standard deviation of 10 years.

(a) On this planet, let $S \in \{0, 20\}$ be cigarettes smoked daily, and let L be life duration. What is the correlation $\rho(S, L) := \text{Cov}(S, L)/\sqrt{\text{Var}(S)\text{Var}(L)}$? Hint: start by using the $\text{Var}(L) = \text{Var}(E[L|S]) + E[\text{Var}(L|S)]$ identity from lecture to get $\text{Var}(L)$. Working out $\text{Var}(S)$ shouldn't be hard. Then attack the two terms of $\text{Cov}(S, L) = E[SL] - E[S]E[L]$. Note that $E[SL] = P\{S = 0\}E[SL|S = 0] + P\{S = 20\}E[SL|S = 20]$.

On Bad Celery Planet, it turns out that (through some poorly understood mechanism) celery is unhealthy. In fact, a single piecce of celery is as unhealthy as a single cigarette on Smoker Planet. However, nobody eats 20 pieces a day for a lifetime. Everybody has a little bit, in varying amounts throughout life. Here is how that works. Each year between age 18 and age 58 a person tosses a fair coin to decide whether to be a celery eater that year. If the coin comes up heads, that person will eat, on average, one piece of celery per day for the entire year (mostly from company vegetable platters). *Given* that one consumes celery for K of the possible 40 years (celery consumption after age 58 has no effect, and everyone lives to be at least 58) one expects to live until age $75 - K/80$, with a standard deviation of 10 years. (So, indeed, eating 1 celery stick per day for the full 40 years is about 1/20 as harmful as smoking 20 cigarettes a day for a lifetime.)

(b) Write L for a person's life duration. On this planet, what is the correlation $\rho[K, L] = \mathrm{Cov}[K, L]/\sqrt{\mathrm{Var}(K)\mathrm{Var}(L)}$? Hint: start by using the $\mathrm{Var}(L) = \mathrm{Var}(E[L|K]) + E[\mathrm{Var}(L|K)]$ identity from lecture to get $\mathrm{Var}(L)$. Working out $\mathrm{Var}(K)$ shouldn't be hard. Then attack $\mathrm{Cov}(K, L) = E[KL] - E[K]E[L]$ as in (a).

Remark: The answer to (b) is much smaller than the answer to (a). So small that it would be *very* hard to demonstrate this effect without a huge sample size. You would need several hundred thousand to be confident that you would see a statistically significant correlation. In the real world, people worry that *many* products have mild carcinogenic effects (effects in the ominous "big enough to matter, small enough to be hard to observe") category. Detectability is a big problem. Moreover, even if you observe the effect in a large sample, people will note that those who eat more celery are statistically different from those who eat less (more health conscious, more prone to eat carrots and ranch dressing, etc.) The effects of these differences could *easily* swamp any effects of the celery itself. One try to "control" for obvious differences (e.g., with multi-variable regressions) but one cannot account for *all* of them, and the question of what to *do* about observed correlation is famously hard. For example, the World Health Organization website says the following about red meat: "Eating red meat has not yet been established as a cause of cancer. However, if the reported associations were proven to be causal, the Global Burden of Disease Project has estimated that diets high in red meat could be responsible for 50,000 cancer deaths per year worldwide. These numbers contrast with about 1 million cancer deaths per year globally due to tobacco smoking, 600,000 per year due to alcohol consumption, and more than 200,000 per year due to air pollution." Shall I eat that burger or not?

D. Let X be a normal random variable with mean μ and variance σ_1^2. Let Y be an independent normal random variable with mean 0 and variance σ_2^2. Write $Z = X + Y$. Let $\tilde{Y}$ be an independent random variable with the same law as Y. Write $\tilde{Z} = X + \tilde{Y}$.

(a) Compute the correlation coefficient ρ of $\tilde{Z}$ and Z.

(b) Compute $E[X|Z]$ and $E[\tilde{Z}|Z]$. Express the answer in a simple form involving ρ. Hint: consider case $\mu = 0$ first and find $f_{X,Z}(x, z)$. You know $F_X(x)$ and $f_{Z|X=x}(z)$. Alternate hint: if X_i are i.i.d. normal with variance σ^2, mean 0, and $n \geq k$ then argue by symmetry that $E[\sum_{i=1}^k X_i | \sum_{i=1}^n X_i = z] = z(k/n)$. Write $X = \sum_{i=1}^k X_i$ and $Y = \sum_{k+1}^n X_i$. Fiddle with k, n, σ^2 to handle the case that σ_1^2/σ_2^2 is rational.

Note that $E[\tilde{Z}|Z]$ is closer to $E[\tilde{Z}] = E[Z]$ than Z is. This is a case of what is called "regression to the mean." Let's tell a few stories about that. An entrant to a free throw shooting competition has a *skill level* that we denote by X, which is randomly distributed as a normal random variable with mean μ and variance 2. During the actual competition, there is an independent *luck factor* that we denote by Y, which is a normal random variable with variance 1 and mean zero. The entrant's overall score is a $Z = X + Y$. If the entrant participates in a second tournament, the new score will be $\tilde{Z} = X + \tilde{Y}$ where $\tilde{Y}$ is an independent luck factor with the same law as Y.

(c) Compute the standard deviation of Z. Given that Z is two standard deviations above its expectation, how many standard deviations above its expectation do we expect $\tilde{Z}$ to be?

Imagine that people in some large group are randomly assigned to teams of 9 people each. Each person's *skill level* is an i.i.d. Gaussian with mean 0 and standard deviation 1. The team's skill level is

the sum of the individual skill levels. You can check that a team's skill level is a Gaussian random variable with mean 0 and standard deviation 3.

(d) Given that a team's total skill level is 6 (two standard deviations above the mean for teams) what do we expect the skill level of a randomly chosen team member to be?

Each drug generated by a lab has an "true effectiveness" which is a normal random variable X with variance 1 and expectation 0. In a statistical trial, there is an independent "due-to-luck effectivness" normal random variable Y with variance 1 and expectation 0, and the "observed effectiveness" is $Z = X + Y$.

(e) If we are *given* that the observed effectiveness is 2, what would we expect the observed effectiveness to be in a second independent study of the same drug?

Remark: The following is from the abstract of the Nosek reproducibility study (recall Problem Set 3) which attempted to reproduce 100 published psychology experiments: "The mean effect size (r) of the replication effects (Mr = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (Mr = 0.403, SD = 0.188), representing a substantial decline." The fact that the effect sizes in the attempted replications were smaller than those in the original studies is not surprising from a *regression to the mean* point of view. Google *Iorns reproducibility* for analogous work on cancer studies.

E. This problem will apply the "regression to the mean" ideas from the previous problem to a toy model for university admissions. Think about admissions at a (somewhat arbitrarily chosen) group of five selective universities: Harvard, Stanford, MIT, Yale and Princeton. For fall 2017, these universities all had (per usnews.com article I looked up) "yield rates" between 65 and 83 percent and class sizes between 1097 and 1703. If we refer to an admission letter to one of these five universities as a *golden ticket* then in all 9841 golden tickets were issued and 7372 were used (i.e., a total of 7372 first year students enrolled at these schools). This means there were 2469 *unused* golden tickets.

Who *had* these 2469 unused golden tickets? Somebody presumbly has a rough answer, but let's just speculate. One wild possibility is that *all* unused tickets were held by 494 lucky students (with 5 unused golden tickets each) who *all* crossed the Atlantic to attend Oxford and Cambridge. If this were true, then *none* of the 7372 golden ticket *users* would have an extra unused ticket. Another wild possibility is that exactly 2469 of the 7372 golden ticket *users* (about 33 percent) have exactly one unused ticket. In any case, the fraction of golden ticket users with an *unused ticket to spare* is between 0 and 33, which implies that the *overwhelming majority* of these 7372 entering students were accepted to the university they attend and to *none* of the other four. The stereotypical "students who apply to all five, get accepted to most" are a small minority. We would need more data to say more than that (where individuals apply, how many apply early admission, how students decide between multiple offers, how admissions criteria vary from place to place, etc.) So let's move to an imaginary (and perhaps not terribly similar) universe where the analysis is simpler. Then we'll do a little math.

In Fancy College Country there are exactly five elite universities and 40,000 elite applicants. All 40,000 applicants apply to all five universities. The *intrinsic strength* of an applicant's case is a normal random variable X with mean 0 and variance 1. When a university reads the application, the university assigns it a *score* $S = X + Y$ where Y is an independent normal random variable with mean

0 and variance 1. Think of X as the college-independent part of an application's strength and Y as the college-dependent part (perhaps reflecting the resonance of the student's background with university-specific goals, as well as the random mood of the admission team). Each student has one value X but gets an independent Y value for each university. Each university admits all applicants with scores above the 95th percentile in score distribution. Since S has variance 2, this means they admit students whose scores exceed $C = \Phi^{-1}(.95) \cdot \sqrt{2} \approx 1.6449 \cdot \sqrt{2} \approx 2.326$ where $\Phi(a) = (2\pi)^{-1/2} \int_{-\infty}^{a} e^{-x^2/2} dx$. To be admitted a student's score must exceed 2.326. Each university expects to admit 5 percent of its applicants.

(a) Compute, as a function of x, the conditional probability that the student is admitted to the first university in the list, *given* that the student's X value is x. In other words, compute the probability that $Y > C - x$.

(b) How large does X have to be for this conditional probability to exceed .05? How about .95? Find the probability that X exceeds the former threshold. And the latter. (Give numerical answers. Note the discrepancy: given their X values, *many* students have a .05 chance, but *very few* have a .95 chance. It is easier to be a contender than a sure thing.)

(c) Let $A(x)$ be the conditional probability that the student is admitted to *at least one* university on the list, given that the student's X value is x. Compute $A(x)$ using Φ and C as defined above.

(d) Argue that the *overall* probability that a student is admitted to at least one university is given by $\int_{-\infty}^{\infty} (1/\sqrt{2\pi}) e^{-x^2/2} A(x) dx$ and that the chance to be rejected by all universities is $\int_{-\infty}^{\infty} (1/\sqrt{2\pi}) e^{-x^2/2} (1 - A(x)) dx$.

(e) Try to compute (d) numerically in a package like wolframalpha and report how it goes. You might (I did) have to fiddle a bit to get it to work. Here's how I did it:

1. To see how wolframalpha represents $\Phi(x)$ type in

   ```
   Integrate[(1/Sqrt[2Pi]) E^(-y^2/2), {y,-Infinity, x}]
   ```

   You get some expression involving erf, which is a close relative of Φ.
2. Click on that to get plaintext. Replace x with $(2.326 - x)$ to get

   ```
   1/2 (1+erf((2.326-x)/sqrt(2)))
   ```

3. Put parentheses about this and raise it to fifth power (to get a wolframalpha friendly expression for conditional chance to be rejected everywhere, given x), multiply by $f_X(x)$ and integrate:

   ```
   Integrate[(1/sqrt(2Pi)) E^(-x^2/2)

   (1/2 (1+erf((2.326-x)/sqrt(2))))^5, {x,-Infinity, Infinity}]
   ```

(f) Briefly justify the following conclusions. Each student has a 0.166363 chance to be accepted at least somewhere. The expected number of students admitted to at least one university is about 6655. The expected class sizes are about 1331 at each school, and each university has a typical yield rate of about .67.

Remark: If admission were completely random (each university takes a student with probability .05 independently of X) then the applicants would have a $1-(.95)^5 \approx .2262$ chance to get accepted to at least one university. We'd expect to see $.2262 \cdot 40000 \approx 9048$ students admited to at least one university, and the yield rate for each university would be roughly .9048. If the selection process were completely determined by X (so that all universities accept exactly the *same* 2000 students) then there would be only 2000 students admitted to at least one university and the yield rate would be .2 (with class sizes of only 400). Our .67 lies between these extremes.

Remark: If some applicants applied to fewer than 5 universities (e.g., due to early admissions) yield rates might be *higher*, since fewer admits would have multiple offers. If the variance of Y were smaller, yields might be *lower* due to greater admission list overlap. If some admits chose *not* to attend one of the 5 schools, that would also decrease yields.

F. The following is one formulation of a famous "two envelope" paradox. Jill is a money-loving individual who, given two options, invariably chooses the one that gives her the most money in expectation. One day Harry, a trusted (and capable of delivering) individual, offers her the following deal as a gift. He will secretely toss a fair coin until the first time that it comes up tails. If there are n heads before the first tails, he will place 10^n dollars in one envelope and 10^{n+1} dollars in the second envelope. (Thus, the probability that one envelope has 10^n dollars and the other has 10^{n+1} dollars is 2^{-n-1} for $n \geq 0$.) Harry will then hand Jill the pair of envelopes (randomly ordered, indistinguishable from the outside) and invite her to choose one. After Jill chooses an envelope she will be allowed to open it. Once she does, she will be allowed to either keep the money in the first envelope or switch to the second envelope and keep whatever is in the second envelope. However, if she decides to switch, she has to pay a one dollar "switching fee."

1. If Jill finds 100 dollars in the first envelope she opens, what is the conditional probability that the other envelope contains 1000 dollars? What is the conditional probability that the other envelope contains 10 dollars?

2. If Jill finds 100 dollars in the first envelope she opens, how much money does Jill expect to win from the game if she does not switch envelopes? (Answer: 100 dollars.) How much does she expect to win (net, after the switching fee) if she *does* switch envelopes?

3. Generalize the answers above to the case that the first envelope contains 10^n dollars (for $n \geq 0$) instead of 100.

4. Jill concludes from the above that, no matter what she finds in the first envelope, she will expect to earn more money if she switches envelopes and pays the one dollar switching fee. This strikes Jill as a bit odd. If she knows she will always switch envelopes, why doesn't she just take the second envelope first and avoid the envelope switching fee? How can she be maximizing her expected wealth if she spends an unnecessary "switching fee" dollar no matter what? How does one resolve this apparent paradox? (Use the hints in the Lecture 24 slides as needed. Even with hints, it may take time to make peace with this problem.)

18.600 Probability and Random Variables

Fall 2019

Central limit theorem and strong law of large numbers

18.600 Problem Set 9

Welcome to your ninth 18.600 problem set! We will explore the central limit theorem and a related statistics problem where one has N i.i.d. samples, one (roughly) knows their standard deviation σ, and one wonders how close the observed average is to the true mean.

The last problem set discussed correlations, including the sort of empirical correlations one observes in real world data. We noted that correlations do not always have clear or simple explanations (like "A causes B" or "B causes A" or "C causes both A and B"). This problem set will explore efforts to understand causation using controlled experiments. According to `https://clinicaltrials.gov/` there are tens of thousands of clinical trials performed every year worldwide. Many have a very simple form: a test group and a control group, and a common variable measured for both groups. Much of what we know about medicine and other areas of science comes from experiments like these.

The idea is that if a variable measured in an experiment has expectation μ and standard deviation σ, then the *average* A of N independent instances of the variable has expectation μ and standard deviation $\overline{\sigma} = \sigma/\sqrt{N}$. If N is large then $\overline{\sigma}$ is small, and A is (by the **central limit theorem**) approximately normal with mean μ and standard deviation $\overline{\sigma}$. This implies $P(|A - \mu| \leq 2\overline{\sigma}) \approx .95$. Since A is close to μ with high probability, it can be seen as an *estimate* for μ. If we can estimate μ accurately, we can *detect* whether μ changes when we modify the experiment. Sampling N independent instances of a random variable (instead of a single instance) is like looking under a $\sqrt{N}$-magnifying microscope. It lets us detect effects that are smaller (by a $\sqrt{N}$ factor) than we could otherwise see.

For example, suppose the amount someone's blood pressure changes from one measurement to another measurement three months later is a random variable X with expectation μ and standard deviation σ. Suppose that if a person is given a blood pressure drug, the change is a random variable $\tilde{X}$ with standard deviation σ and expectation $\mu - \sigma$.

If you try the drug on *one* person and blood pressure decreases, you can't tell if this is due to the drug or chance. But consider $A = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\tilde{A} = \frac{1}{N}\sum_{i=1}^{N} \tilde{X}_i$ where X_i are independent instances of X and $\tilde{X}_i$ are independent instances of $\tilde{X}$. Now A and $\tilde{A}$ are roughly *normal* with standard deviation $\sigma/\sqrt{N}$ and means μ and $\mu - \sigma$. If $N = 100$, then $E([\tilde{A} - A)] = -10\overline{\sigma}$, which is (in magnitude) *ten times* the standard deviation of A and thus $10/\sqrt{2} \approx 7$ times the standard deviation of $(\tilde{A} - A)$. This is now a "visible" difference.

In statistics, one defines a *p-value* to be the probabilty that an effect as large as the one observed would be obtained under a "null hypothesis." In the trial described above, one might assume as a null hypothesis that A and $\tilde{A}$ are identically distributed (and roughly normal) with standard deviation $\overline{\sigma}$. Then *experimentally observe* $x = (\tilde{A} - A)$. The p-value is $\Phi\big(x/(\overline{\sigma}\sqrt{2})\big)$, which is the probability that $(\tilde{A} - A) \leq x$ *under the null hypothesis*. One (arguably unfortunate) convention is to say x is *statistically significant* if $p \leq .05$ (or $p \leq .025 \approx \Phi(-2)$, which roughly means that either $x \leq -2\text{SD}(A - \tilde{A})$ or $x \geq 2\text{SD}(A - \tilde{A})$). The problem with the convention is that given many trials, each measuring many things, one sees many "significant" results due to chance. It can be hard to explain to the layperson that "statistically significant" is not a synonym for "meaningful". In some settings, one expects *most* statistically significant results to be due to chance, not an underlying effect.[1]

[1] In the discussion above, we assume that the standard deviations of X and $\tilde{X}$ are both roughly equal to a known value σ. If σ is not known, we can replace it with an *approximation* s (called a *sample standard deviation*) computed from the data itself. When A is a sample mean, the number of standard deviations (of A) by which it exceeds its null hypothesis value is sometimes called a z**-score**. A t**-score** is the same except that the standard deviation of A is estimated using s in place of σ. If you want to know the probability that a t-score is large, you have to consider that one way for it to be large is if the z-score is large, but another is if s happens by chance to be much less than σ. Google *Student's t-test* or *Welch's t-test* or *two-sample t-test* to find out how to compute p-values that take both of these things into account. These tests are based on the assumption that *either* X and $\tilde{X}$ are normal *or* the sample size is large enough so that the sample means are roughly normal (and the sample standard deviation is not too likely to be unusually small). We won't say any more about t-tests in the course, but you'll see them a lot if you read academic papers, and it's good to know what they're talking about. (The chocolate study mentioned above uses a t-test.)

If you google *Bohannon chocolate* you can read an entertaining exposé of the willingness of some journals to publish (and news organizations to publicize) dubious statistically significant results. Bohannon conducted a tiny ($N = 15$) trial, tested many parameters, and *happened* to find $p < .05$ for one of them. The trial was real, but anyone familiar with basic statistics who read the paper would be almost certain that the finding ("dark chocolate causes weight loss") was due to chance. (Also too good to be true.) It was widely reported anyway.

A stricter "5 sigma standard," common in physics, requires $|x| \geq 5\text{SD}(\tilde{A} - A)$, or $p \leq \Phi(-5) \approx .0000003$. The recent *Higgs boson* discovery used that standard. *Very* roughly speaking, you smash tiny things together lots of times and measure the released energy; if you get more measurements in the *Higgs boson* range than you expect due to chance (and the result is significant at the 5 sigma level) you have observed the particle.

Before launching an experiment, you should have a common sense idea of what the magnitude of the effect might be, and make sure that N is large enough for the effect to be visible. For example, suppose you think babies who watch your educational baby videos weekly will grow up to have SAT scores a 10th of a standard deviation higher than babies who don't. Then first of all, you should realize that this would be a pretty big effect. (If 12 years of expensive private schooling/tutoring raise SAT score one standard deviation—perhaps a high estimate—your videos would have to do more than an average *year* of expensive private schooling/tutoring.) And second, you should realize that even if the effect is as big as you think, you can't reliably recognize it with a trial involving 100 babies. With 10,000 babies in a test group and 10,000 in a control group, the effect would be clear. But can you realistically conduct a study this large?

A. **Remark:** To prepare for the next problem, suppose that you discover a market inefficiency in the form of a mispriced asset. Precisely, you discover an asset priced at \$10 that has a $p > 1/2$ chance to go up to \$11 over the next day or so (before reaching \$9) and a $(1 - p) < 1/2$ chance to go down to \$9 (before reaching \$11). By buying r shares at \$10 and them selling when the price reaches \$9 or \$11, you have an opportunity to make a bet that will win r dollars with probability $p > 1/2$ and lose r dollars with probability $(1 - p)$. Let's ignore transaction costs and bid-ask spread. (And assume that, unlike all those people who merely *think* they can recognize market inefficiencies, you *actually can*. Assume also that your wisdom was obtained legally — so no risk of an insider trading conviction!) So now you effectively have an opportunity to bet r dollars on a p coin with $p > 1/2$. The question is this: how much should you bet? In expectation you will make $pr + (1 - p)(-r) = (2p - 1)r$ dollars off this bet, so to maximize your expected payoff, you should bet *as much as you possibly can*. But is that really wise? If you repeatedly bet all our money on p-coins, it might not be long before you lose everything. The *Kelly strategy* (which comes from assuming utility is a logarithmic function of wealth — look it up) states that instead of betting everything, you should bet a $2p - 1$ fraction of your current fortune. The next problem is a simple question about this strategy.

1. Problem 67: Consider a gambler who, at each gamble, either wins or loses her bet with respective probabilities p and $1 - p$. A popular gambling system knkown as the Kelly strategy is to always bet the fraction $2p - 1$ of your current fortune when $p > 1/2$. Compute the expected fortune after n gambles of a gambler who starts with x units and employs the Kelly strategy.

B. If is N is (approximately) a normal random variable with mean μ and variance σ^2, then this problem will refer to the interval $[\mu - 2\sigma, \mu + 2\sigma]$ as the 95-*percent interval* for N. The random variable N lies within this interval about a $\Phi(2) - \Phi(-2) \approx .95$ fraction of the time. We'd be surprised if N were *far* outside this interval. On the other hand, one can show that $1.5 \leq |N| \leq 2.5$ about 12 percent of the time: hence, outcomes *near the edge* of this interval are *not surprising at all*. Give the 95-percent interval (whose endpoints are mean plus or minus two standard deviations) for each of the quantities below. Try to solve these problems quickly and in your head if you can (okay to write interval without showing work). The better you get at this, the more you'll apply it in daily life. (Simple rule: when you sum N i.i.d. copies of something, SD is

multipled by $\sqrt{N}$. When you average N i.i.d. copies, SD is divided by $\sqrt{N}$. Remember that SD is $\sqrt{npq}$ for binomial and $\sqrt{\lambda}$ for Poisson.)

1. A university admits 600 students and expects 60 percent to accept their offers. Give the 95-percent interval for the university's yield rate.

2. 10000 people are infected with a certain flu virus. The expected duration of symptoms is 10 days, with standard deviation 5 days. Give the 95-percent interval for average duration.

3. There is a group of 100 college students at an elite university. After ten years, the income of each student will be an independent random variable with mean \$150,000 and standard deviation \$50,000. Give the 95-percent interval for the overall average income of the student collection.

4. Lisa the Lyft driver gets an independent rating from each passenger. The scores are 5 with probability .8 and 4 with probability .2 so her expected rating is 4.8. Give a 95-percent interval for her average rating after 100 trips. Laura the Lyft driver's scores are 5 with probability .9 and 1 with probability .1 so her expected rating is 4.6. Give the 95-percent interval for her average after 100 trips. **Hint:** Laura's scores should fluctuate a lot more than Lisa's.

5. Alice takes a course with 2 midterms (each 25% of grade) and one final (50% percent of grade). Partial credit rules vary, but *roughly speaking* each midterm has 25 key ideas (and the final 50 key ideas) that one either gets or doesn't. Alice gets each of these (independently) with probability .8. Compute the 95-percent interval for her overall percentage.

6. Bob's favorite basketball team scores $X_1 + 2X_2 + 3X_3$ points in a game, where X_i are independent Poissons with $\lambda_1 = 15$, $\lambda_2 = 30$, $\lambda_3 = 10$. Give the 95-percent interval for the score.

7. Carol's fund makes 25 risky (independent) investments per year. Each earns a return with expectation 5 percent and SD 20 percent. Give the 95-percent interval for the *average* return.

C. In the modifications below the (roughly) normal random variable N (for which you gave a 95 percent interval) is replaced by a (roughly) normal $\tilde{N}$ with different mean but (roughly) same standard deviation. Indicate the *number of standard deviations (of N)* by which the mean is shifted. That is, compute $(E[\tilde{N}] - E[N])/SD(N)$. (Okay to give number without showing work.) This describes how *detectable* the change is. (The corresponding 95-percent intervals overlap if this number is less than 4; see `http://rpsychologist.com/d3/cohend/` for vizualization.) And whether one can say, "Given independent instances of N and $\tilde{N}$, the latter will be noticeably better with high probability."

1'. The university offers a nicer financial aid package, increasing expected yield to 66%.

2'. The patients take antiviral drugs that reduce expected duration from 10 to 9 days.

3'. The group of students takes 18.600, which makes them more savvy and productive by every measure—and in particular increases their expected income by \$5000.

4'. Both Lyft drivers begin offering free bottled water, which raises their expected scores by .08.

5'. Alice stops studying altogether, which reduces her correctness probability from .8 to .08.

6'. Bob switches allegiance to the Golden State Warriors, who (at that moment) average 120 points per game.

7'. Carol hires a smarter quantitative analyst and increases expected returns to 7 percent.

Remark: Does it disturb anyone that an effect as large as the one I snarkily attribute to 18.600 is still too small too to reliably measure, even with an $N = 100$ randomized study?

Remark: Lyft computes a driver score based on the past 100 rides, and Uber computes a score based on the past 500 rides. Both companies encourage drivers to maintain scores above 4.8. (Lower scores bring "needs improvement" flags from Lyft and disqualify drivers from UberBLACK in New York. Accounts may be deactivated if scores go *too* much lower.) If you peruse message boards, you'll see that drivers (and passengers) worry a lot about the extent to which scores fluctuate due to chance. It is somehow disconcerting that a single 4 is no big deal (some riders consider 4 a good score) but a 4 from half your riders can get you fired. Alice finds it similarly disconcerting that even with study her 95-percent interval may span two or three letter grades. Grades are noisy measurements, maybe more so than we would like. Actual NBA scores are roughly normal with mean above 100, SD about 12.
`https://squared2020.com/2015/11/01/hypothesis-testing-is-nba-scoring-up-this-year/`
Here unpredictability is part of the appeal—better teams don't *always* win. Note: shot clocks and court-crossing times might make "time between shots" follow a non-exponential distribution—and may cause "number of shots taken" to vary less than a Poisson of the same mean, at least outside of the game's final minute. See `https://moldham74.github.io/AussieCAS/papers/Gon.pdf`. Also, if possessions alternate, then "possessions per team" are not independent for two opposing teams, and one has to account for this to model win probabilities. And as long as our problem takes place in an imaginary universe, let's say the Celtics are the ones with 120 points per game. :) See `https://www.teamrankings.com/nba/stat/points-per-game` for current stats.

Remark: Here is another model for Alice's grade: suppose Alice has an ability level $x \in [0, 100]$, and each problem has a difficulty level $y \in [0, 100]$, and Alice solves the problem with probability 1 if $x > y$ and with probability 0 otherwise. If the exam problems have difficulties $1, 2, \ldots, 100$ then Alice's score is the integer part of x with probability one. Unlike the model above, this one predicts that repeated tests yield the same score. (This can be checked empirically; google *inter-rater reliability*.) In reality, even with careful design, it is not possible to make an exam perfectly reliable in this way. (A test that just measures one's height in centimeters would be *nearly* perfectly reliable but would still have some measurement error.)

Remark: See `https://www.act.org/content/dam/act/unsecured/documents/Research-Letter-about-ACT-Writing.pdf` for an ACT reliability study (from when essay had a 36-pt scale). It writes:

1. *A score of 20 on the ACT composite would indicate that there is a two-out-of-three chance that the student's true score would be between 19 and 21.*
2. *A score of 20 on ACT math, English, reading or science would indicate that there is a two-out-of-three chance that the student's true score would be between 18 and 22.*
3. *A score of 20 on ACT wrting would indicate that there is a two-out-of-three chance that the student's true score would be between 16 and 24.*

The writing score is especially noisy. Roughly doubling interval width to get a 95 percent interval, one might phrase it this way: *A score of 28 on writing would indicate a 95 percent chance that the student's true score would be between 20 (below average at most colleges) and 36 (best possible).* Some argue that even noisy measurements are informative, and should be used but given low weight—just as though the essay were one of many exam problems. (Recall the previous problem set settings, where the noisier a measurement is, the less one changes conditional expectation in response to it.) Others argue that noisiness causes stress and all but forces students to take exams multiple times. Other measurements (interviews, letters, scores based on extracurriculars, etc.) might be just as noisy, but the noise may be harder to quantify. Last I checked, MIT does not require ACT or SAT essays.

D. On Blueberry Planet, researchers plan to assemble two groups with N people each. Each group will take a fitness test before and after a six month period. Let A_1 be the average fitness

improvement for the control group and A_2 the average fitness improvement for a group assigned to eat blueberries. The improvement for each *individual* in the control group is an independent random variable with variance σ^2 and mean μ. The improvement for each individual in the blueberry eating group is an independent random variable with variance σ^2 and mean $\mu + rb$ where r is an unknown parameter and b is the number of blueberries the blueberry eaters are assigned to eat each day. (We are assuming a *linear dose response* so 2 blueberries have twice the effect of 1 blueberry, etc.) Assume N is large enough so that A_1 and A_2 are approximately normal with given means and variances. Suppose that there is a limited research budget for blueberries, so Nb is fixed. For the purpose of estimating the size of r, would it be better to take N large and b small, or to take N small and b large? Explain.

Remark: Realistically, the linearity of the dose response probably only holds up to a certain point, and there is some practical upper bound on b. Also, it is unlikely that blueberries would really be the most expensive part of this experiment. But if one replaces "blueberries" with years of exposure to a new educational technique (which requires training teachers, etc.) or a new crime prevention technique, it might make sense to assume Nb is limited.

Remark: Drug abuse programs like DARE would be worth their cost even if they only saved a few people. But it is hard to say (google "is DARE effective") how measurable the effects are. Could it be that (like so many things educators and parents do...) it has a long term effect that is large enough to matter but too small to reliably detect with the experiments we can do?

E. Interpret/justify the following: the *p-value* computed from a simple experiment (as described in the intro to this pset) is a random variable. If an effect size is large enough so that the *median* p-value is $\Phi(-2)$ then in a similar trial with 6.25 times as many participants the *median* p-value would be $\Phi(-5)$.

Remark: In a previous problem set, we discussed Cautious Science Planet and Speculative Science Planet, where hypotheses with different *a priori* likelihood were tested. Another way two planets could differ is in the p-value they use to define significance. Should medicine and other sciences should adopt the 5σ standard used in physics (and somehow assemble the resources to make their data sets 6.25 times larger) or maybe an even stricter standard? This would lead to a much smaller number of positive findings, but the findings would be more trustworthy. On the other hand, if you google *Is the FDA too conservative or too aggressive?* you can read an argument by an MIT professor and student that the FDA should approve drugs for incurable cancers (when the patient will die anyway without treatment) using a *lower* standard of evidence than they currently use. A more general question (does exercise alleviate depression?) might be addressed using *many* kinds of experiments. Some argue that many small experiments are more informative than one large one, since the idiosyncracies of the experiment designs average out; but *meta-analysis* (combining multiple studies to get a conclusion) is a tricky art, and there may be a lot of bias in what is and isn't published.

F. Kevin and James are playing a simplified two-person version of Jeopardy (no daily doubles or final jeopardy) with exactly 60 questions, and James is the stronger player. Let J_i be 1 if James answers the ith question right, -1 if he answers it wrong, zero otherwise. Let K_i be 1 if Kevin answers the ith question right, -1 if he answers it wrong, zero otherwise. Write $\alpha_i = J_i - K_i$. Assume that the α_i are i.i.d. with mean .3 and variance 1. The total game score difference (final score for James minus final score for Kevin) is $D = \sum \alpha_i \beta_i$ where β_i is the "value" of the ith question. James wins if D is positive.

1. Compute the mean, variance and standard deviation of D if each β_i is equal to 900.

2. Compute the mean, variance and standard deviation of D if there are six β_i values for each number in the sequence $\{200, 400, 600, 800, 1000, 400, 800, 1200, 1600, 2000\}$. (Note, since 400 and 800 appear twice, this means 12 questions of each of those values.) Hint: put something like

```
6 (Sum[(200 i)^2, {i, 1, 5}] + Sum[(400 i)^2, {i, 1, 5}] )
```

into wolframalpha.com as part of your variance calculation

3. Assuming D is roughly normal (with mean/variance you just computed) approximate the probability that Kevin wins in each of the two scenarios above. Work out numerical answers.

Remark: Sometimes things that make games fun to watch (like point values that differ from question to question) also make outcomes less predictable. That is, they increase the probability that a "stronger" player will lose. James Holzhauer recently won 32 Jeopardy games in row and Ken Jennings won 74. Would you expect *even longer* streaks if the questions were assigned equal values and there were no daily doubles or final jeopardy wagers? Note: in the above story James expects 18 more questions (net) than Kevin, a huge difference. But both Holzhauer and Jennings averaged about 35 correct answers (net) per game, per some site I just googled, with opponent average likely in the single digits. Of course, for the purpose of maintaining a streak, strength of average opponent might matter less than strength of occasional exceptional opponent.

G. Harry knows that either Hypothesis X is true, and a test will give a positive answer 80 percent of time, or Hypothesis X is false, and a test will give a positive answer 5 percent of the time. Harry thinks *a priori* that Hypothesis X is equally likely to be true or false. Harry does his own test and the result is positive.

(a) Given that the test is positive, what is Harry's revised assessment of the probability that Hypothesis X is true?

Sherry also thinks *a priori* that Hypothesis X is equally likely to be true or false. Sherry knows (from her research world connections) that exactly ten groups (including Harry's) have conducted independent tests of the kind that Harry conducted. She knows that they have all had ample time to publish the results, but she has not yet heard the results. Sherry has electronic access to the prestigious *We Only Publish Positive and Original Results Medical Journal* (WOPPORMJ). Sherry knows that each group with a positive test would immediately submit to WOPPORMJ, which would publish only the first one received. So WOPPORMJ will have a publication if and only if *at least one* of the tests was positive. Sherry opens WOPPORMJ and finds an article (by Harry) announcing the positive result.

(b) Given Sherry's new information (that *at least one* of the ten tests was positive), what is Sherry's revised assessment of the probability that Hypothesis X is true?

That evening, Sherry and Harry meet for the first time at a party. They discuss their revised probability estimates. Harry tells Sherry that he is upset that she has not raised her probability estimate as much as he has. They decide to try to come up with a revised probability using all of the information they have together. The conversation starts like this:

1. **Harry:** I computed my probability with correct probabilistic reasoning. Then you came along and said you knew that nine other teams tested for X, but you don't know *anything* about what they found. You have given me no new information about Hypothesis X and thus no reason to change my assessment of the probability it is true.

2. **Sherry:** I computed my probability with correct probabilistic reasoning. When I did my computation, I knew that WOPPORMJ had accepted a paper by someone named Harry. I have learned nothing by meeting you and see no reason to change my view.

But, being smart and curious people, they continue to talk and reason together.

(c) Assuming that they both apply sound logic, what happens? Do they end up both agreeing with Sherry's probability estimate, or both agreeing with Harry's estimate, or both agreeing on something else, or continuing to disagree in some way? (There is a hint on the next page, but don't look at it before you need to.)

Remark: Some people think that *all* experimental data should be published—regardless of whether it is negative or unoriginal (and also regardless of whether it is bad for the financial bottom line or the political agenda of the group funding the study...) Look up "clinical trials registry" to read about relevant efforts this direction.

HINT BELOW:

HINT: You actually need another assumption to pin down the answer. First solve the problem with the first assumption below (which may be what you were tacitly assuming anyway). Then solve it with the second assumption, which will give you a different answer.

1. The order in which the ten groups completed their tests was *a priori* random (all 10! permutations equally likely and independent of hypothesis truthfulness and test outcomes). So to describe a state space element, one needs to know the truthfulness of Hypothesis X (two possibilities), the outcomes of the 10 tests (2^{10} possibilities), and the submission order (10! possibilities). So $|S| = 2 \cdot 2^{10} \cdot 10!$. Harry had no idea before submitting his paper what the ordering was, and Harry and Sherry have no further information about that (beyond the fact that they know Harry's paper was accepted).

2. Harry knows that his was the first group to complete the test.

18.600 Probability and Random Variables

Fall 2019

Entropy, Martingales and Finance

18.600 Problem Set 10

Welcome to your tenth 18.600 problem set! We'll be thinking a bit about the efficient market hypothesis, risk neutral probability, martingales, and the optional stopping theorem. These ideas are commonly applied to financial markets and prediction markets, but they come up in many other settings as well. Indeed, if X is any random variable with finite expectation, then as one observes more and more information, one's revised conditional expectation for X evolves as a martingale, and the optional stopping theorem applies to the sequence of revised expectations.

Suppose that you know you have an exactly $2^{-10} = 1/1024$ chance of dying during the next 12 months. (You can see at `https://www.ssa.gov/oact/STATS/table4c6.html` what fraction of US men and women your age die during a given year; the 1/1024 figure is way too high for women but actually a little low for men.) Now which would you prefer?

1. If you die, it happens a random time during the year (no knowledge in advance).
2. You witness the outcome of one coin toss every month over the course of ten months, and you die at the end of the year if all are heads.
3. All coins are tossed at once at the end of the year, and you die if they are all heads.

If you choose the second option, your conditional probability of dying that year will evolve as a martingale (starting at 1/1024, then jumping to 0 or 1/512, then jumping to 0 or 1/256, etc.) If you choose the third option, there is a single jump (from 1/1024 to either 1 or 0) that happens all at once. This choice posed here may seem morbid, but in fact real life poses frequent analogs of this question (involving early cancer diagnoses, mammograms, genetic tests, etc.) and the answers are not easy. How much information about our revised chances do we really want to have? How do we respond emotionally to martingale ups and downs? How do we react when the first six tosses are heads one year, and the future is suddenly scarier?

Any fan of action movies knows that the heroes frequently face circumstances where (based on non-movie logic) the conditional probability that they survive the adventure appears very low. (*C3PO: Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to 1.*) This (sequentially revised) conditional probability is like a martingale that gets very close to zero, then somehow comes back to a moderate value, then gets close to zero again, then returns to moderate, then gets *extremely close* to zero in a big climactic scene, and then somehow gets back to one. This behavior is unlikely for actual martingales. But nobody argues that the stories that get made into movies (fictional or otherwise) are typical. We like to tell the stories with close calls and happy endings, even if most stories aren't like that.

This problem set also features problems about *entropy* (an extremely important concept in (for example) statistical physics, information theory, and data compression) as well as Markov chains, which play an important role in operations research, computer science, and many other areas. If you google *seven shuffles* you'll find a famous result about how many shuffles are required to adequately mix up a deck of cards: in this case, a "shuffle" is a step in a Markov chain on the set of 52! permutations of a standard deck of cards.

A. FROM TEXTBOOK CHAPTER NINE:

1. Problem/Theoretical Exercises 7: A transition matrix is said to be doubly stochastic if $\sum_{i=0}^{M} P_{ij} = 1$ for all states $j = 0, 1, \ldots, M$. Show that if such a Markov chain is ergodic, with $(\pi_0, \pi_1, \ldots, \pi_M)$ the stationary row vector, then $\pi_j = 1/(M+1), \quad j = 0, 1, \ldots, M$.

2. Problem/Theoretical Exercises 9: Suppose that whether it rains tomorrow depends on past weather conditions only through the last 2 days. Specifically, suppose that if it has rained yesterday and today, then it will rain tomorrow with probability .8; if it rained yesterday but not today, then it will rain tomorrow with probability .3; if it rained today but not yesterday, then it will rain tomorrow with probability .4; and if it has not rained either yesterday or today, then it will rain tomorrow with probability .2. Over the long term, what proportion of days does it rain?

3. Problem/Theoretical Exercise 13: Prove that if X can take on any of n possible values with respective probabilities $P_1, \ldots, P_n$, then $H(X)$ is maximized when $P_i = 1/n, i = 1, \ldots, n$. What is $H(X)$ equal to in this case?

4. Problem/Theoretical Exercise 17: Show that, for any discrete random variable X and function f,
$$H(f(X)) \leq H(X).$$

B. A standard die is repeatedly rolled until the first time it comes up 6. Let X be the full sequence of outcomes *up to that point*. For example, maybe $X = \{1, 4, 2, 3, 2, 5, 6\}$ or $X = \{3, 4, 2, 6\}$ or $X = \{6\}$. Compute the entropy $H(X)$. **Hint:** Let K be the length of X. Argue that $H(X) = H(X, K)$. Compute the entropy of K directly (it is a geometric random variable). Then recall the identity $H(X, K) = H(K) + H_K(X)$ from the lecture slides and find $H_K(X)$.

C. **Relative entropy and world view:** Suppose that there are n possible outcomes of an athletic tournament. I assign probabilities $p_1, p_2, \ldots, p_n$ to these outcomes and you assign probabilities $q_1, q_2, \ldots, q_n$ to the same outcomes. If the ith outcome occurs and $p_i > q_i$ then I will interpret this as evidence that my probability estimates are better than yours, and that perhaps I am smarter than you. In short, I will feel smug. Suppose that my precise smugness level in this situation is $\log(p_i/q_i)$. Then before the event occurs, my *expected* smugness level is $\sum p_i \log(p_i/q_i)$.

(a) Show that my *expected* smugness level is always non-negative, and that it is zero if and only if $p_i = q_i$ for all i. (Hint: use some calculus to find the vector $(q_1, q_2, \ldots, q_n)$ that minimizes my expected smugness level. This is rather like Problem A.1 above.)

(b) Suppose that if outcome i occurs, your smugness level is $\log(q_i/p_i)$, so that your expected smugness level is $\sum q_i \log(q_i/p_i)$. We agree *a priori* that our combined smugness level will be zero no matter what (your smugness is by definition negative one times my smugness). However, you expect your smugness level to be positive (and mine to be negative) while I expect my smugness level to be positive (and yours to be negative). Try your best to give a short intuitive explanation for why that is the case.

(c) Look up the term *relative entropy* and explain what it has to do expected smugness.

Remark: I expect an *infinite* amount of smugness if I assign positive probability to things that you assign zero probability. We sometimes say our probability distributions are *singular* when this is the case. As a practical matter in politics, it might be a bad thing if my professed probability distribution is close to singular with respect to yours on some of the great unknowns (e.g., likelihood that certain tax cuts help the economy or that certain public investments provide net benefits or that some religious or philosophical ideas are true). The discrepancies might make it hard for us to find *any* common political ground, even if we are both utilitarians seeking the greater good. On the other hand, discrepancies are betting/trading opportunities. If our probability differences are real (and not political smokescreen) perhaps we can make a policy bet in the form of a policy that funds your priorities if your predictions pan out and my priorities if my predictions pan out.

Remark: The *Smugness Game* is played as follows. You specify a vector q, I specify p, and when the ith outcome occurs you pay me $\log(p_i/q_i)$, which is equivalent to me paying you $\log(q_i/p_i)$. You can check that no matter what p I choose, if you know the "true probability vector" you maximize your expected payout by choosing that for q. Both players are incentivized to give their best probability assessments. Some people would like to see weather prediction and political prediction teams challenge each other to play this game.

D. Solve the following problems:

1. Suppose Harriet has 30 dollars. Her plan is to make one dollar bets on fair coin tosses until her wealth reaches either 0 or 80, and then to go home. What is the expected amount of money that Harriet will have when she goes home? What is the probability that she will "win," i.e., that she will have 80 when she goes home?

2. Harriet uses the phrase "I think A" in a precise way. It means "the probability of the event A, conditioned on what I know now, is at least .5". Which of the following is true:

   (a) Harriet thinks she will lose.

   (b) Harriet thinks that there will be a time when she thinks she will win.

   (c) Harriet thinks that her amount of money will reach 41 and subsequently reach 20 before the game is over.

3. What is the expected number of bets she will make before reaching 0 or 80? **Hint:** Let $F(n)$ be the expected number there would be if she started with n dollars for $n \in \{0, 1, 2, \ldots, 80\}$. So $F(0) = F(80) = 0$. Use a conditional expectation argument (involving value of first toss) to show $F(k) = 1 + \frac{F(k+1)+F(k-1)}{2}$, which implies $[F(k+1) - F(k)] - [F(k) - F(k-1)] = -2$. Guess a quadratic function F might be and show it is the only function with these properties.

E. Complete the derivation of the Black-Scholes formula for European call options, as outlined in the lecture slides, by explicitly computing

$$E[g(e^N)]e^{-rT},$$

where N is a normal random variable with mean $\mu = \log X_0 + (r - \sigma^2/2)T$ and $g(x) = \max\{0, x - K\}$.

F. David Aldous of UC Berkeley devised and told me about the following problem. How many people in a given US presidential election cycle do we expect to have their probability of becoming president *at some point* exceed 10 percent? In other words, if we look at those prediction market charts (and pretend the plots are true continuous martingales) how many candidates do we *expect* will have their number at some point exceed 10 percent? Let's consider two ways to approach this problem. (**Note:** if a *continuous* martingale like Brownian motion is below .1 at one time .1 and above .1 at a later time, then it must pass through .1 at some intermediate time. If you want to avoid thinking about continuous-time martingales, just consider a discrete-time martingale with tiny increments that has this property.)

(a) Assume that at some point (well before the election) every person in the world has some small probability p to become president. The ith person has probability p_i and $\sum p_i = 1$. Assume that the ith person's probability evolves as a continuous martingale; then it has a $p_i/.1 = 10p_i$ chance to reach ten percent in the prediction markets at some time. Sum over i to get an overall expected number of people who reach this threshold.

(b) Imagine a gambler who adopts the following strategy. Whenever a candidate reaches 10 percent, the gambler buys a contract on that candidate for \$10 (which will pay \$100 if the candidate wins) and holds it until the end of the election. Then at the end of the election the gambler is certain to receive \$100 (since the gambler will have purchased a contract on the winner at the first time the winner's price reached \$10). Argue that by the optional stopping theorem, the gambler makes zero money in expectation. So the expected amount of money spent on contracts must be \$100.

Here is a variant. Suppose there is a .85 chance you will get married eventually (at least once). Imagine that aliens watching your life from afar are placing bets on who your first spouse will be, and that contract prices evolve as continuous martingales. Call somebody an "almost first spouse" if at some point the market probability that this person is your first spouse exceeds .5.

(c) Assuming continuity of martingales, how many almost first spouses do you expect to have over your lifetime?

(d) Call a person a "serious contender" if their probability exceeds .1 at some point. How many serious contenders do you expect to have?

Remark: Does it seem a little strange that somebody who doesn't know your romantic history at all could make what appear to be substantive probabilistic assessments about your future love life? Also, did I need to bring aliens into this story? Could it be your friends or parents, or maybe you yourself, assessing these probabilities?

Remark: I have no data on this, but I have heard it argued that people in relationships are irrational, because their subjective relationship-viability estimates fluctuate more than martingale theory predicts. If your estimate of the probability you will marry your current significant other regularly alternates between over .8 and below .2, and you expect this to happen several more times, then your evolving subjective probability estimates will not (according your current subjective probability) evolve as a martingale. This puts you in violation of economic rationality assumptions.

Assuming rationality/consistency, the probability measure *you* assign to *your own* future revised-probability trajectory should make it a martingale. A one-step example of a violation: "Today I think there is an 80 percent chance we'll marry some time in 2018, but my *expectation* of what my probability will be after tomorrow's date is 90 percent." Of course, none of us is *actually* fully rational in the sense of having consistent and well defined probabilities for all future outcomes.

G. Imagine that at this particular moment on the currency market, one dollar has the same value as one euro. Let R be the event that (at some time during the next year) the euro rises in value relative to the dollar, so that the euro becomes worth two dollars. Let P_d be the cost, in dollars, of a contract that gives you one dollar if and when the event R occurs. Let P_e be the cost, in euros, of a contract that gives you one euro if and when event R occurs. Argue that one should expect $P_e = 2P_d$.

Remark: Assuming no interest, you can interpret P_d as the risk neutral probability of R (using dollars as the *numéraire*), but you can also interpret P_e as the risk neutral probability of R (using euros as the *numéraire*). This should convince you that the risk neutral probability of R cannot *always* be interpreted as a "general consensus of the subjective probability that R will occur". (That interpretation is only reasonably if the value of money does not depend on whether R occurs.) In October of 2016, currency traders believed (correctly, it turns out) that the price of the Mexican peso versus the US dollar would fall significantly if Trump were elected. Based on this, the risk neutral probability of Trump's election should have been *lower* with pesos as the numéraire than with dollars as the numéraire. (Imagine the extreme case: if it were known that Trump's election would make pesos worthless, then the price — in pesos — of a contract paying a peso if Trump were elected would be zero.)

Concluding remark: Congratulations on finishing (or at least reading to the end of) your final problem set! Your problem sets and the remarks therein have *briefly* introduced you to many topics: Powerball odds, Occam's razor, hypothesis testing, the Doomsday argument, p-values, Siegel's paradox, subprime lending, modern portfolio theory, diversification, the capital asset pricing model, idiosyncratic versus systemic risk, utility and risk aversion, Poisson bus inefficiency, Gompertz mortality, radioactive decay and half life, Cohen's d, clinical trials, open primary voting, Pascal's wager, infinite expectation paradoxes, correlation versus causation, the Kelly strategy, least squares regression, regression to the mean, publication bias, relative entropy, the Black-Scholes derivation, and the dependence of risk neutral probability on the numéraire.

Take a moment to review the problem sets and recall the stories you have forgotten. Topics that appear only in problem sets (not in lecture notes or practice exams) are not likely to be on your final. I nonetheless hope you review and retain at least some understanding of these stories. You could not have understood them without the math you have learned, and reviewing them may help you solidify your math skills and integrate probability into your thinking about the world. I also hope some of you use probability to solve big problems: to fundamentally improve our approach to science, medicine, criminal justice, economics, engineering, and so forth. Failing that, I hope you'll at least have fun with all of this.

Best of luck!

Okay, one more (optional) paradoxical remark: The following is known in philosophy circles as the *Sleeping Beauty problem*. On Sunday night a woman agrees to the following experiment. She will be put into a medically induced sleep. Then a fair coin will be tossed. If the toss comes up heads, she will be woken and interviewed *both Monday and Tuesday* mornings before being returned to sleep (each time with an amnesia-inducing drug that makes her forget the experience). If the coin comes up tails, she will be woken and interviewed *only on Monday* morning. Either way, she awakes without interview on Wednesday and the experiment is over.

When she is interviewed (on either Monday or Tuesday), she does not know either what day it is or what the coin outcome was. So we can think of $\{Mon, Tue\} \times \{H, T\}$ as the sample space corresponding to her uncertainty at that point. Basic symmetries suggest that during any interview she interprets the states $\{Mon, H\}$ and $\{Tue, H\}$ as being equally likely, and also that she interprets and $\{Mon, H\}$ and $\{Mon, T\}$ as being equally likely; hence she sees each as having probability 1/3. (She knows that $\{Tue, T\}$ isn't possible if she is awake.)

But here's the paradoxical part. On Sunday night, she thinks there is a 1/2 chance that the coin is heads, but she also *knows* that on Monday morning, according to her updated probability estimate, she will think there is a 2/3 chance. At first glance this seems to violate the fact that sequentially revised conditional expectations evolve as martingales. How can she rationally think there is a 1/2 chance now but also *know* that she will think there is a 2/3 chance tomorrow morning?

Mathematically, the problem is that we cannot say she is just revising her probability by conditioning on new information. She is also *forgetting* and/or *losing track of time* in a way that creates a new kind of uncertainty. The day of the week is not an unknown from the woman's point of view on Sunday, but it *becomes* unknown as of Monday morning. There have been over a hundred philosophy written papers about this problem, some of which reach different conclusions about the Monday morning heads probability or "credence." See `https://www.quantamagazine.org/solution-sleeping-beautys-dilemma-20160129/` or look up the survey by the mathematician Peter Winkler. (You can also google "sleeping beauty problem" and read the Wikipedia artice.) For a simpler "forgetfulness" story, imagine you know that a coin toss just came up heads, but you also know that a year from now you'll have forgotten this fact, so that at that point you'll subjectively think the chance the coin was heads is .5. Your sequentially revised heads probability cannot be a martingale in this story, since it violates the optional stopping theorem. Forgetting is problematic.

So... try not to forget too much of what you learned in this course. And best of luck again!

18.600 Probability and Random Variables

Fall 2019