Cauchy, beta, gamma and the infinite expectation paradox

18.600 Problem Set 7, due April 13

Welcome to your seventh 18.600 problem set! This problem set features problems about beta, Γ, 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 standard interpretations.

Many of you are familiar with *Pascal's wager*. The general 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.

Please stop by my weekly office hours (2-249, Wednesday 3 to 5) for discussion.

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. FROM TEXTBOOK CHAPTER SIX: Theoretical Exercise 12: Show that the jointly continuous (discrete) random variables $X_1, \ldots X_n$ are independent if and only if their joint probability density (mass) function $f(x_1, \ldots, x_n)$ can be written as

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g_i(x_i),$$

for nonnegative functions $g_i(x)$, $i = 1, \ldots, n$.

C. Let p be the fraction of MIT students who love Taylor Swift — or, more precisely, the fraction who will *say* they love Taylor Swift 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 Taylor Swift, 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 Taylor Swift. 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.

D. 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 amount of money is in the second envelope. However, if she decides to switch envelopes, she has to pay a one dollar "switching fee."

(a) 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?

(b) 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?

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

(d) 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?

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. (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.

Remark: On Divorce Planet, each person has an inborn and immutable quantity p chosen uniformly from $[0, 1]$. Upon reaching adulthood, each person marries somebody with essentially the same p value (assortative mating). With probability p the marriage lasts forever; otherwise it ends within a few years, the individuals remarry (again, somebody with a similar p value), and the experiment starts over. The story is similar to the IVF model discussed above, but with IVF cycles replaced by weddings and "conceiving a child" replaced by "entering a lasting marriage" and the initial p distribution being uniform. Using the analysis in the above problem, one can show that on Divorce Planet, 1/2 of first marriages, 2/3 of second marriages, and 3/4 of third marriages end in divorce. Our world may be very different from Divorce Planet, but according to `https://www.psychologytoday.com/blog/the-intelligent-divorce/201202/the-high-failure-rate-second-and-third-marriages` the divorce rates in the modern U.S. are strikingly similar: 50 percent for first marriages, 67 percent for second marriages, and 73 percent for third marriages. The linked to article speculates about several possible reasons the rate might be higher for later marriages, without really considering the one that applies on Divorce Planet (i.e., that there are persistent attributes that make some people more prone to divorce than others, and that one expects people entering later marriages *on average* to have more such attributes than people entering first marriages). I certainly do not claim to know which explanations are correct on our planet.

F. Suppose $X_1, X_2, \ldots, X_6$ are independent Cauchy random variables. Compute the probability that $X_1 + X_2 + X_3 > X_4 + X_5 + X_6 + 3$. (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. Let X be a Γ random variable with parameters $\lambda = 1$ and $\alpha = n$ where n is an integer. Let Y be an exponential random variable with parameter $\lambda = 1$. Derive the variance for X from the variance for Y using a "waiting time until nth bus" story.

H. 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."

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 one medicine cures your headache with probability .5, one with probability .6, and you don't know which is which. Or maybe one airline has good food with probability .5 and another 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 grading: How meaningful are small differences in GPA? And does it make a difference whether a university reports only a letter grade score as opposed to having a continuum of score options? To think about this, consider two stories. At Normal University, each student taking a class gets real number for a grade. One student's grades are independent normal random variables with mean 0 and variance 1. Another student's are the same but with mean .25. If both students take a class, then the difference between the (slightly) stronger student's score and the other student's score has variance 2 and mean .25. If the students take 32 courses, then the average difference has mean .25 and variance $2/32 = 1/16$, hence standard deviation .25. Thus the average score difference is one standard deviation about zero, which means that it will be positive with probability about 5/6 (using rule of thumb that standard normal is one standard deviation below the mean about 1/6 of the time). So there is about a 5/6 chance the stronger student has a higher GPA.

At A Or B University the raw grades are the same as above, but the transcript simply records an A if the raw grade is positive and a B if it is negative. So one student has an A with probability .5 and the other has an A with probability about .6 (since a normal is less than a quarter standard deviation above its mean about 60 percent of the time). On the 4 point scale ($A = 4$ and $B = 3$) one student expects a 3.5 and one expects a 3.6. The GPA difference (after N classes are taken) has standard deviation $.7\sqrt{N}$ and expecation .1. Thus, after 49 classes, the chance that the stronger student has the higher GPA is about the same as it woud be at Normal University after 32 classes. Generally, the decision to report only the binary data (A or B) means that it takes about 50 percent longer to distinguish between the two students (at any specified accuracy level).

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?