Cauchy, beta, and gamma random variables
18.600 Problem Set 7, due November 3

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. **BETA APP:** Textbook Chapter 5, 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. **BAYESIAN COIN:** 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 $P\big(p \leq a \big| X_1 = H\big)$. 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. **EMPIRICAL LOVE:** Let p be the fraction of MIT students who are happy to see Taylor Swift and Travis Kelce together — or, more precisely, the fraction who will *say* they are when you ask (making it clear that you *absolutely* require a yes or no answer). Let's make believe that your initial Bayesian prior for p is uniform on $[0, 1]$. Now ask three fellow students (actually do this!) one at a time if they are happy with Swift/Kelce 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 is happy with Swift/Kelce. 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, Marvel Movies, Lin-Manuel Miranda, Tom Brady and Ariana Grande. If you have ideas for next year let me know. :) The ideal is someone/something well known but whose popularity within MIT we would be *a priori* unsure of, so that the uniform prior doesn't seem *too* unreasonable. (I have seen enough opinion polls to suspect very few national *politicians* are loved by more than 70 percent of us...)

D. **GAMMA MOMENTS:** 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. **IVF:** 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.

1. 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).

2. Compute the conditional probability that the couple conceives during the kth cycle, given that they did not conceive during the first $(k-1)$ cycles, using the following approach. 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.]

3. 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 the density function of a beta random variable (for some a and b) 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 3 (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 3. 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. **CAUCHY MEAN:** Suppose $X_1, X_2, \ldots, X_{12}$ are independent Cauchy random variables. Compute the probability that $\sum_{i=1}^{6} X_i < 12 + \sum_{j=7}^{12} X_j$. (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.)

Solve one of G1 and G2 (your choice). If you solve both, we'll take your higher score.

G1. **ACCOUNTING FOR TASTE:** 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 $\frac{(f+1)}{(f+1)+(r+1)}$. 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 13 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.

G2. **REVISED ELECTORAL COUNTRY:** Consider a variant of Simple Electoral Country where state vote percentages still range roughly from 25 to 75 but spacings are *random* and *irregular*. Assume $n = 51$ states of equal size. Let X_i be the vote percentage of the ith state. Assume the X_i are *i.i.d. and uniform* on $[25, 75]$. Such X_i tend to spread "roughly evenly" on $[25, 75]$.

First query: in this Revised Electoral Country, how often is the electoral college winner (the one winning majorities in at least 26 states) the same as the popular vote winner? In 10 million Mathematica simulations, I found a discrepancy (popular vote winner losing electoral college) in a .16478 fraction of cases. This is close to 1/6 (and gets even closer when $n >> 51$). Let's see why.

1. Let A_i be i.i.d. random variables taking values in $\{-12.5, 12.5\}$ with .5 probability each. Let B_i be uniform random variables (independent of the A_i and each other) on $[-12.5, 12.5]$. Write $X_i := 50 + A_i + B_i$ and check that the X_i thus defined are i.i.d. and uniform on $[25, 75]$.
2. Show that $\text{Var}[A_i] = 3\text{Var}[B_i]$.
3. Write $A = \sum A_i$ and $B = \sum B_i$ and $X = \sum X_i$. Central limit theorem (coming later) says A, B roughly normal (mean 0, given variance). Discrepancy occurs if $|B| > |A|$ and A, B have opposite sign. Show this happens about 1/6 of time. **Hint:** Set $\tilde{B} := \sqrt{3}B$. $\text{Var}(\tilde{B}) = \text{Var}(A)$ so $(A, \tilde{B})$ has near rotation symmetry. Think about slope-$\sqrt{3}$ lines and 60-degree angles...

Remark: Okay, that's the 1/6. What else can we compute?

4. `(1/51)Sum[2(51 choose j)(1/2)^51((j-24) choose 2)*25/(j+1),{j,26,51}]`

   can be typed into wolframalpha and the result is .149538. I claim that this computes the expected number of overall percentage points one needs to flip to swing the election. If you agree, can you justify this? **Hint:** use what you know about beta random variables to say that if j random variables are uniform and independent in $[0, 25]$, then the expected value of the kth lowest is $k \cdot 25/(j+1)$. Note also that $\sum_{k=1}^{m} k = \binom{m+1}{2}$.

5. `Sqrt[1/51]Sqrt[2500/12]Integrate[(1/Sqrt[2Pi])E^-(x^2/2)|2x|,{x,-Infty,Infty}]`

   yields 3.22526 and is meant to estimate the expected discrepancy in popular vote percentage points between the candidates. Can you justify this? **Hint:** observe that $\text{Var}(X_1) = 2500/12$ and $\text{Var}(\frac{X}{51}) = \frac{1}{51^2}\text{Var}(X) = \frac{1}{51^2} \cdot 51 \cdot \text{Var}(X_1) = \frac{1}{51} \cdot 2500/12$ so $\text{SD}(\frac{X}{51}) = \sqrt{1/51} \cdot \sqrt{2500/12}$.

Remark: The model in this problem predicts elections where (a) the popular vote gap is about 3 percentage points on average, (b) the "number of vote flips needed to change outcome" tends to be much smaller than that, and (c) the popular vote winner loses the electoral college in one of six elections. Among the last dozen or so US elections (with two popular/electoral discrepancies, lots of close races) does this match some of what we see empirically? Less so for earlier US history?

Remark: The answer in Part 4, while small, is larger than in the Pset 5 version. Intuitive reason? Well, in the current problem, big electoral college victories happen (one side winning 29, 30, 31, 32 or more states) and in these scenarios it takes lots of vote flips to change the outcome. But big electoral college wins may not correspond to *such* big popular wins. For example, *given* that a party wins 30 states its conditional popular vote expectation is only $(30 * 62.5 + 21 * 37.5)/51 \approx 52.2$ percent.