Exponentials and normal approximations

18.600 Problem Set 6, due April 7

Welcome to your sixth 18.600 problem set! 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. Please stop by my weekly office hours (2-249, Wednesday 3 to 5) for discussion.

A. FROM TEXTBOOK CHAPTER FIVE:

1. Problem 23: One thousand independent rolls of a fair die will be made. Compute an approximation to the probability that the number 6 will appear between 150 and 200 times inclusively. If the number 6 appears exactly 200 times, find the probability that the number 5 will appear less than 150 times.

2. Problem 32: The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = 1/2$. What is

   (a) the probability that a repair time exceeds 2 hours?

   (b) the conditional probability that a repair takes at least 10 hours, given that its duration exceeds 9 hours?

3. Theoretical Exercise 9: If X is an exponential random variable with parameter λ, and $c > 0$, show that cX is exponential with parameter λ/c.

4. Theoretical Exercise 29: Let X be a continuous random variable having cumulative distribution function F. Define the random variable Y by $Y = F(X)$. Show that Y is uniformly distributed over $(0, 1)$.

5. 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 you how f_X is stretched, squashed, and translated when you replace X by $aX + b$) it makes it easier to remember a couple of the formulas on the story sheet. The third problem above is a special case of the last one.

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.

(a) Compute the expectation of the time T_n at which the number of bacteria reaches n.

(b) Compute the variance of T_n.

(c) Are both of the answers above unbounded, as functions of n? Give a rough numerical estimate of the values when $n = 10^{50}$.

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:

(a) 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 $\tilde{X}$.

(b) 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.

(c) 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.

(a) 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).

(b) 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.

(c) 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).

(d) 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?)