Exponentials and normal approximations

18.600 Problem Set 6, due April 1

Welcome to your sixth 18.600 problem set! Let’s warm up by thinking about data analysis. 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 \( \sqrt{npq} = \sqrt{40 \cdot 0.8 \cdot 0.2} = \sqrt{6.4} \approx 2.52 \), 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. Please stop by my weekly
office hours (2-249, Wednesday 3 to 5) for discussion.

A. GOOD MEASURE:

1. Textbook Chapter 5, Theoretical Exercise 9: If \( X \) is an exponential random variable with
parameter \( \lambda \), and \( c > 0 \), show that \( cX \) is exponential with parameter \( \lambda/c \).

2. Textbook Chapter 5, Theoretical Exercise 30: Let \( X \) have probability density \( f_X \). Find the
probability density function of the random variable \( Y \) defined by \( Y = aX + b \).

3. Explain, in a few sentences, the argument that the circle is the disjoint union of countably many
sets that are all rotations of each other, as in the third to last of the Lecture 17 slides. (You can
also read the Wikipedia article on Vitali sets and/or the axiom of choice.)

REMARK: If you internalize the idea of the second problem (you understand 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 first problem above is a special case of the second one.

B. PETRI DISH: 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 $n = 10^{50}$.

**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. **EMPIRICAL FAIR:** 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 confidentially 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 $d = (E[\tilde{X}] - E[X]) / SD[X] = .016$. 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. SAVING DEMOCRACY: 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 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 (i.e., the amount your vote changes \( E[W] \), compared to your not voting at all) is

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

where \( p_i \) is the probability that \( B_i \) wins the second primary \textit{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. COFFEE FOR CLOSERS:** 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 agents, it takes years to confidently determine who is better. If you think you can tell after a few weeks, you are deluding yourself—the noise to signal ratio is too high. This problem appeared (without the real estate 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 probability .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?