Central limit theorem and strong law of large numbers

18.600 Problem Set 9, due April 29

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. **KELLY STRATEGY (FROM TEXTBOOK):** 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. **CONFIDENCE INTERVALS IN YOUR HEAD:** 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. **DETECTABILITY:** 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 Milwaukee Bucks, who average about 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: When I first wrote this problem, Lyft computed a driver score based on the past 100 rides, and Uber computed a score based on the past 500 rides. Both companies encouraged drivers to maintain scores above 4.8. (Lower scores brought "needs improvement" flags from Lyft and disqualified drivers from UberBLACK in New York. Accounts would be deactivated if scores got *much* lower.) Many drivers (and passengers) worried a lot about chance fluctuations. It is disconcerting that a single 4 is no big deal (some riders consider 4 a good score) but a 4 from half your riders gets 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 was 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. **STUDY DESIGN WITH RESOURCE SCARCITY:** 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 empirically*) 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? If the effect is not empirically detectable, can we be sure it is positive and not negative? This is not just an issue for DARE. Try googling *is diversity training effective empirically* or *is anti-harassment training effective empirically*. There is a lot of literature on these complex and challenging topics.

E. **STUDY DESIGN WHEN YOU WANT A LOW p-VALUE:** 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. **PUBLICATION BIAS:** 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 ON NEXT PAGE

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.