Markov chains, central limit theorem, strong law of large numbers

18.600 Problem Set 9, due May 5

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 about 37,000 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 that 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 small 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.

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?

This problem set will also say a bit about Markov chains, which play an important role in operations research, computer science, and many other areas. If you google *seven shuffles* you'll discover a famous mathematical result about how many shuffles are required to adequately mix up a deck of cards: in this case, a "shuffle" is understood as a step in a Markov chain on the set of 52! permuations of a standard deck of cards.

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

A. FROM TEXTBOOK CHAPTER NINE:

1. Problem/Theoretical Exercises 7: A transition matrix is said to be doubly stochastic if $\sum_{i=0}^{M} P_{ij} = 1$ for all states $j = 0, 1, \ldots, M$. Show that if such a Markov chain is ergodic, with $(\pi_0, \pi_1, \ldots, \pi_M)$ the stationary row vector, then $\pi_j = 1/(M+1), \quad j = 0, 1, \ldots, M$.

2. Problem/Theoretical Exercises 9: Suppose that whether it rains tomorrow depends on past weather conditions only through the last 2 days. Specifically, suppose that if it has rained

yesterday and today, then it will rain tomorrow with probability .8; if it rained yesterday but not today, then it will rain tomorrow with probability .3; if it rained today but not yesterday, then it will rain tomorrow with probability .4; and if it has not rained either yesterday or today, then it will rain tomorrow with probability .2. Over the long term, what proportion of days does it rain?

B. 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, but we wouldn't be so suprised if N were near the edge of this interval. 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 also that SD is $\sqrt{npq}$ for binomial and $\sqrt{\lambda}$ for Poisson.)

1. 100 students will take the ACT exam. In this population, each student's score is an independent random variable with mean 20 and standard deviation 5. Give the 95-percent interval for the average score.

2. An Uber driver gets an independent random rating from each passanger: these ratings have expectation 4.65 and standard deviation .8. Give a 95-percent interval for the average rating over 400 trips.

3. There is a group 100 of 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. The number of customers who show up to a restaurant over a given day is Poisson with parameter $\lambda = 400$. Give a 95-percent interval for that number.

5. A university admits 600 students and expects 60 percent to accept their offers. Give the 95-percent interval for the university's yield rate.

6. 10000 people are infected with a certain flu virus. The expected duration of symptoms is 12 days, with standard deviation 5 days. Give the 95-percent interval for average duration.

C. 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 100 students take a test-prep course that increases expected ACT score by 2.

2'. The Uber driver begins offering passengers free bottled water, which raises the expected scores from 4.65 to 4.73.

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'. The restaurant launches an advertising campaign that increases λ to 600.

5'. The university offers a nicer financial aid package, increasing expected yield to 70%.

6'. The patients take antiviral drugs that reduce expected duration from 12 to 11 days.

Remark: Last I read (google this), Uber drivers have to keep their average (over last 500 rides) above 4.6. Apparently most riders give 5 stars for a "normal ride" but some give 3, so chance fluctuations affect drivers at the margin. In principle, Uber could reduce the chance component by subtracting from each rider's score the average score that rider gives to other drivers — so a 3 from somebody who always gives 3 would count the same as a 5 from somebody who always gives 5. But this would take some power away from the customers. (Maybe 5-givers *like* helping all their drivers a bit. Maybe 3-givers *want* fewer drivers to qualify. Is not clear whether riders know how ratings are used.) If you believe the numbers at `http://www.businessinsider.com/leaked-charts-show-how-ubers-driver-rating-system-works-2015-2` you can try to assess how "at risk" the bulk of drivers (in the 4.7 to 4.8 range) are of falling below the threshold due to chance.

Remark: Recall also the regression problems on the last problem set, which addressed how to modify one's conditional expectation of a "true value" given a noisy measurement of that value. (The noisier you believe the measurement to be, the less you change your conditional expectation in response to the measurement.) Note also that in reality that there may be other sources of systemic noise (weather affecting restaurant traffic or flu symptoms, actions of competing universities affecting yield rates, etc.) beyond what we have accounted for.

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?

D. 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”) 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?

E. Interpret and 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. 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?

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.