Prof. David Draper
Department of Statistics
University of California, Santa Cruz

AMS 131: Take-Home Test 3 *[520 total points]*

Due date: upload to `canvas.ucsc.edu` by **11.59pm Sun 1 Sep 2019**

1. *[130 total points]* (biology) *Limnology* is the study of inland waters (both saline and fresh), including their biological, chemical and hydrological properties. One common outcome variable in studies in this branch of biology is pH, because the acidity of a lake can be an important factor in determining the abundance of fish and other wildlife living in and near it. According to the web site `www.lenntech.com/aquatic/acids-alkalis.htm`,

> Unpolluted deposition (or rain), in balance with atmospheric carbon dioxide, has a pH of 5.6. Almost everywhere in the world the pH of rain is lower than this. The main pollutants responsible for acid deposition (or acid rain) are sulfur dioxide (SO_2) and nitrogen oxides (NO_x). Acid deposition influences mainly the pH of freshwater. ... Most freshwater lakes, streams, and ponds have a natural pH in the range of 6 to 8. Acid deposition has many harmful ecological effects when the pH of most aquatic systems falls below 6 and especially below 5. Here are some effects of increased acidity on aquatic systems:
>
> - As the pH approaches 5, non-desirable species of plankton and mosses may begin to invade, and populations of fish such as small-mouth bass disappear.
> - Below a pH of 5, fish populations begin to disappear, the bottom is covered with undecayed material, and mosses may dominate near-shore areas.
> - Below a pH of 4.5, the water is essentially devoid of fish.

You're a limnologist out in the field studying a lake — sufficiently remote that you had to backpack in to get to it — and this lake looks like it may already have been damaged by acid rain. The only pH measurement kit you could bring with you in your backpack is rather crude: it's known to give unbiased pH measurements that fluctuate around the true value with an SD of 0.15 and an approximately Normal distribution for its measurement errors. You'll be surveying enough lakes on this trip that you can't bring water samples back with you; you need to estimate their pH values in the field.

You're wondering if the pH of the lake you're now standing in front of is below 5; let's agree to call any such lake *threatened*. You decide to take one or more pH measurements to reduce your uncertainty about the lake's status.

This problem is about *measurement error*, so I need to introduce some notation and concepts. Before you've measured anything, let Y_i be a random variable capturing the uncertainty in your prediction of observation i, as i runs from 1 to n. In words, the standard measurement error model encourages you to additively decompose Y_i into the sum of (the true quantity being measured) plus (systematic error, also known as *bias*) + (random error):

$$(\text{observation})_i = (\text{truth}) + (\text{bias}) + (\text{random error})_i \,. \tag{1}$$

This model requires an act of imagination to formulate, because the only thing we get to observe (the number on the left side of the equation) is broken into the sum of three things we can't observe; you may therefore wonder at its usefulness, but (as we'll see) it's actually quite helpful.

Let θ stand for the true value of the thing being measured (in this problem, θ is the true pH of the lake); let b stand for the bias in the measurement process; and let the e_i be the random measurement errors. Then symbolically equation (1) looks like

$$\begin{aligned} Y_1 &= \theta + b + e_1 \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \\ Y_n &= \theta + b + e_n \,. \end{aligned} \tag{2}$$

In the standard measurement error model, the e_i are regarded as IID random variables (this assumption is only reasonable if (i) the measurements are performed in a logically independent manner and (ii) you try hard to ensure that each observation is performed in precisely the same way) with mean 0 (any mean other than 0 gets absorbed into the bias term) and finite standard deviation σ. Define $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{e}_n = \frac{1}{n}\sum_{i=1}^n e_i$.

(a) Show that $\bar{Y}_n = \theta + b + \bar{e}_n$; show that $V(\bar{e}_n) = \frac{\sigma^2}{n}$; and therefore show that $E\big(\bar{Y}_n\big) = \theta + b$ and $V\big(\bar{Y}_n\big) = \frac{\sigma^2}{n}$. Intuitively, why is the variance of $\bar{e}_n$ smaller than the variance of any of the e_i going into $\bar{e}_n$? Show that your results in this part of the problem imply that $\bar{Y}_n$ only converges in probability to the truth θ if $b = 0$. Show that the typical amount $\mathrm{RMSE}\big(\bar{Y}_n\big) = \sqrt{E\left[\big(\bar{Y}_n - \theta\big)^2\right]}$ by which $\bar{Y}_n$ is likely to differ from θ (RMSE stands for *root mean squared error*) is given by the Pythagorean expression

$$\mathrm{RMSE}\left(\bar{Y}_n\right) = \sqrt{\frac{\sigma^2}{n} + b^2}\,, \tag{3}$$

and that therefore this also only goes to 0 as more data accumulates if $b = 0$. *[80 points]*

Suppose for the rest of this problem that the true pH of this lake is 5.1, so that in fact it's not actually threatened.

(b) If you take only a single water sample and process it with your pH kit, what's the probability that you'll incorrectly conclude that this lake is threatened? Show your work. *[10 points]*

(c) You're not happy with the misclassification probability in (b), and you decide to remedy this by taking $n > 1$ independent water samples from the lake and basing your assessment on their mean pH value $\bar{Y}_n$. How large does n need to be to make the probability of {incorrectly concluding that this lake is threatened} 0.5% or less? Be explicit about all aspects of your probability model, including all of the assumptions you make and whether you think they're reasonable. *[40 points]*

2. *[140 total points]* (medicine) Hypertension is a medical condition in which a person's blood pressure is chronically elevated. (A reminder: blood pressure is measured with two numbers called *systolic* (higher) and *diastolic* (lower), in a deeply anachronistic scale called mmHg (millimeters of mercury); blood pressures are stored as data in the form "systolic over diastolic" [i.e., 115 over 75 or 115/75;

Subject	1	2	3	4	5	6	7	8	9	10	11	12	Mean	SD
Before	200	174	198	170	179	182	193	209	185	155	169	210	185.3	17.1
After	191	170	177	167	159	151	176	183	159	145	146	177	166.8	14.9
Difference	+9	+4	+21	+3	+20	+31	+17	+26	+26	+10	+23	+33	18.6	10.1

Table 1: *Before and after results for $n = 12$ hypertensive patients treated with Captopril.*

and ideal blood pressures range from 90/60 to 120/80.) Persistent hypertension is one of the risk factors for strokes, heart attacks, heart failure and arterial aneurysm, and is a leading cause of chronic renal failure; as of 1999, it was estimated that 29% of American adults were hypertensive. A U.S. public health goal in 2000 was to lower this rate to 16% by 2010, but thing have actually gotten worse since then: the *American Heart Association* estimated in 2018 that *46%* of all U.S. adults are hypertensive (although part of the increase is due to a change in the definition of high blood pressure from (above 140 systolic) to (above 130 systolic)). Diet and exercise can go a long way to lower blood pressure, but drugs are also sometimes needed (particularly given how hard it is to get Americans to exercise and eat in a healthier way :-) .

The online reference *Wikipedia* notes that "*Captopril* is an angiotensin-converting enzyme (ACE) inhibitor used for the treatment of hypertension and some types of congestive heart failure. Captopril was the first ACE inhibitor developed and was considered a breakthrough both because of its novel mechanism of action and also because of the revolutionary development process. ... The development of Captopril was among the earliest successes of the revolutionary concept of *structure-based drug design*. The renin-angiotensin-aldosterone system (a hormone system that helps regulate long-term blood pressure and blood volume in the body) had been extensively studied in the mid-20th century, and it had been decided that this system presented several opportune targets in the development of novel treatments for hypertension."

Captopril was developed in the mid 1970s; MacGregor et al. (1979, *British Medical Journal*) published the results of a clinical trial on its effects. Systolic blood pressures (in mmHg) were measured for $n = 12$ representatively-chosen hypertensive patients, before and after taking Captopril for a long enough time period for the drug to work. Before any data had been gathered, let (B_i, A_i) be a pair of random variables signifying the before and after blood pressure readings for person i in the study (as i runs from 1 to n), and define $D_i = (B_i - A_i)$ and $\bar{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i$; the realized values of these random variables are given in Table 1.

(a) Estimate the average effect Δ of Captopril in the population to which you believe it's appropriate to generalize here, and explicitly identify that population. Is this estimated effect large in clinical terms? Attach a standard error to your estimated effect, and construct an approximate 95% confidence interval for Δ, explicitly identifying all assumptions you're making. Is the estimated effect statistically significant? What do you conclude about Captopril's usefulness in treating hypertension? Explain briefly. *[80 points]*

(b) Figure 1 presents the scatterplot matrix for the before and after systolic blood pressure readings on these patients and the differences, with pairwise correlations noted.

(i) The experimental setup used by the investigators in this problem is called a *repeated-measures* design leading to a *paired comparison*, because blood pressure was measured twice on the same n people and the analysis focused on the differences (before – after). Another way the experiment could have been run — this is called a *completely randomized* design — would be to (I) choose $2\,n$ hypertensive people in a representative manner and (II) randomize n of them to receive a placebo (the control group) and the other n to

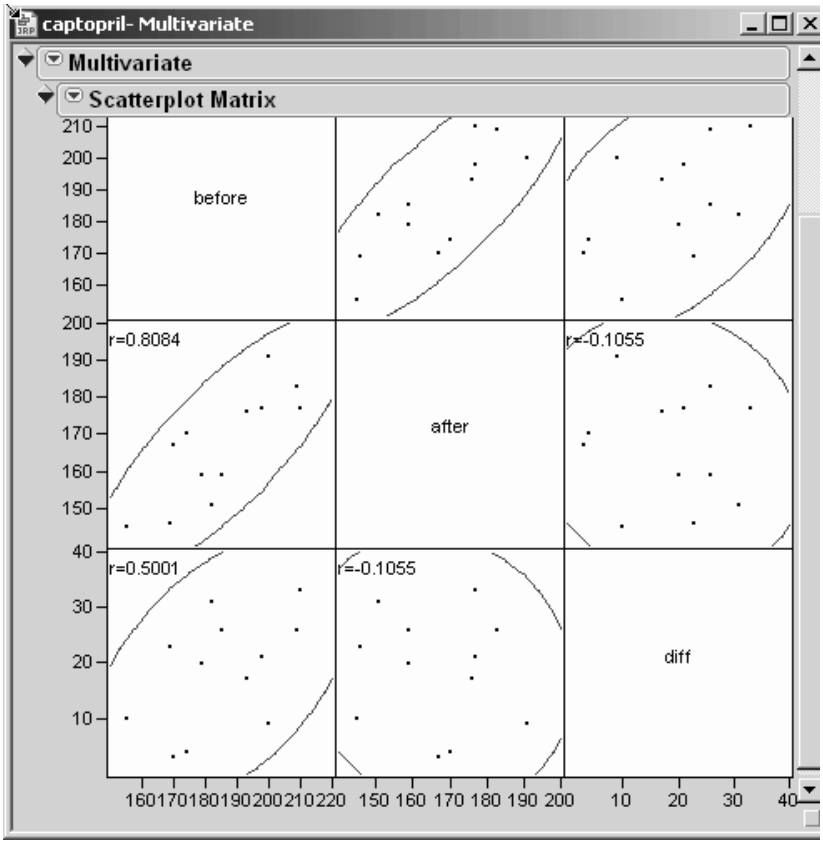

Figure 1: *Scatterplot matrix for the variables* `before`, `after`, *and* `diff`.

receive Captopril (the treatment group). The realized (B_i, A_i) values in Table 1 can be used to make a good guess at what the data set would have looked like if the investigators had used a completely randomized design instead of their paired comparison: the only difference would be that the B_i and A_i values in Table 1 would have been independent, because the data values in column i of the table would have come from two different people.

The estimate of the treatment effect with the completely randomized design would have been $\hat{\Delta} = \bar{B}_n - \bar{A}_n$, where $\bar{B}_n = \frac{1}{n}\sum_{i=1}^{n} B_i$ and $\bar{A}_n = \frac{1}{n}\sum_{i=1}^{n} A_i$, but notice that this is the same as $\bar{D}_n = \frac{1}{n}\sum_{i=1}^{n}(B_i - A_i) = \left(\frac{1}{n}\sum_{i=1}^{n} B_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} A_i\right)$. Let V_{RM} and V_{CR} denote the variance of $\bar{D}_n$ under the repeated-measures and completely-randomized designs, respectively; also let σ_B^2 and σ_A^2 denote the population variances of B_i and A_i, respectively, and define $\rho \triangleq \rho(B_i, A_i)$. Show that

$$V_{RM}\left(\bar{D}_n\right) = \frac{\sigma_A^2 + \sigma_B^2 - 2\,\rho\,\sigma_A\,\sigma_B}{n} \quad \text{and} \quad V_{CR}\left(\bar{D}_n\right) = \frac{\sigma_A^2 + \sigma_B^2}{n}\,, \tag{4}$$

and that therefore the *efficiency* of the RM design when compared with CR is given by

$$e(RM, CR) \triangleq \frac{V_{CR}\left(\bar{D}_n\right)}{V_{RM}\left(\bar{D}_n\right)} = \frac{\sigma_A^2 + \sigma_B^2}{\sigma_A^2 + \sigma_B^2 - 2\,\rho\,\sigma_A\,\sigma_B}\,. \tag{5}$$

Show, using the data values in Table 1 and the correlations in Figure 1, that in this experiment RM was 5.0 times more efficient than CR (and that doesn't even reflect the fact that CR used $2\,n$ patients instead of the n patients in RM). *[40 points]*

(ii) Does the effect of the drug seem to be constant across the 12 patients, or is there a tendency for the drug to have a larger or smaller effect for people whose initial blood pressure was high than for those whose initial reading was lower? Which (if any) of the correlations in Figure 1 supports this conclusion? Explain briefly. *[20 points]*

3. *[130 total points]* (binomial and negative binomial sampling) You and I are both getting ready to sample from a Bernoulli process with unknown success probability $0 < \theta < 1$. You decide to use *binomial sampling*: you propose to

(1) set a fixed known number $n \geq 1$ of Bernoulli trials in advance,

(2) observe that many trials, and

(3) record the random number S of successes you see.

I instead propose to use *negative binomial sampling*: I'll watch the same process that you do, but I'll

(1′) set a fixed known number $s \geq 1$ of successes in advance,

(2′) observe the Bernoulli trials until I've seen s successes, and

(3′) record the random number N of trials that were needed to get that many successes.

Question, to be answered by parts (a–c) of this problem below: if your S equals my s and my N equals your n, should you and I draw essentially the same conclusions about θ?

(a) Briefly explain why your probability model for S should be Binomial(n, θ), so that your S has PMF

$$f_S(s \,|\, n, \theta) = \left(\begin{array}{c} n \\ s \end{array} \right) \theta^s (1-\theta)^{n-s} \, I_{\{0,1,\ldots,n\}}(s) \,, \tag{6}$$

and why a natural estimator of θ for you to use is therefore $\hat{\theta}_B = \frac{S}{n}$. Show that $E\left(\hat{\theta}_B\right) = \theta$, so that $\hat{\theta}_B$ is unbiased; show further that $SE\left(\hat{\theta}_B\right) \triangleq \sqrt{V\left(\hat{\theta}_B\right)} = \sqrt{\frac{\theta(1-\theta)}{n}}$; and briefly explain under what conditions the distribution of $\hat{\theta}_B$ should be approximately Normal. *[50 points]*

(b) Recall that if X records the number of failures before the sth success, then $X \sim$ Negative Binomial(s, θ), with PMF

$$f_X(x \,|\, s, \theta) = \left(\begin{array}{c} s+x-1 \\ x \end{array} \right) \theta^s (1-\theta)^x \, I_{\{0,1,\ldots\}}(x) \,. \tag{7}$$

(i) Briefly explain why the random N I'll observe with my sampling method is related to X via the simple expression $N = X + s$. *[10 points]*

(ii) Show that the PMF of N is

$$f_N(n \,|\, s, \theta) = \left(\begin{array}{c} n-1 \\ s-1 \end{array} \right) \theta^s (1-\theta)^{n-s} \, I_{\{s,s+1,\ldots\}}(n) \tag{8}$$

(*Hint:* use Theorem 1.8.3 of DS (page 34): for all integers $n \geq 1$ and all integers $k = 0, 1, \ldots, n$, $\left(\begin{array}{c} n \\ k \end{array} \right) = \left(\begin{array}{c} n \\ n-k \end{array} \right)$) *[10 points].*

Notice how similar equations (6) and (8) are; this encourages the idea that you and I will get more or less the same answers about θ if I use the estimator $\hat{\theta}_{NB} = \frac{s}{N}$.

(iii) Use results from class or DS about $E(X)$ and $V(X)$ to show that $E(N) = \frac{s}{\theta}$ and $V(N) = \frac{s(1-\theta)}{\theta^2}$ *[10 points]*. Then use the Delta Method with your results about N to show that $E\left(\hat{\theta}_{NB}\right) \doteq \theta$, so that $\hat{\theta}_{NB}$ is approximately unbiased, and that $SE\left(\hat{\theta}_{NB}\right) \triangleq \sqrt{V\left(\hat{\theta}_{NB}\right)} \doteq \sqrt{\frac{\theta(1-\theta)}{E(N)}}$ *[20 points]*.

(iv) Use Jensen's Inequality to show that — in a refinement to the Delta Method — $E\left(\hat{\theta}_{NB}\right) > \theta$, so that $\hat{\theta}_{NB}$ is actually biased on the high side. It can be shown (you're not asked to show this) that $E\left(\frac{s-1}{N-1}\right) = \theta$ (call this fact $(*)$); for a fixed observed value n of N, use $(*)$ to show that the bias of $\hat{\theta}_{NB}$ goes to 0 like $\frac{1}{n}$, so that — for large N — $\hat{\theta}_{NB}$ is indeed approximately unbiased. *[20 points]*

(c) Looking at the expressions for the means and standard errors (SEs) of $\hat{\theta}_B$ and $\hat{\theta}_{NB}$, is it true that you and I will come to pretty much the same conclusions about θ with our different but related sampling methods? Explain briefly. *[10 points]*

4. *[120 total points]* (public health) In one of the largest human experiments ever conducted, in 1954 a randomized controlled trial was run to see whether a vaccine developed by a doctor named Jonas Salk was effective in preventing paralytic polio. A total of 401,974 children (ages 6–9), chosen to be representative of those who might be susceptible to the disease, were randomized to two groups: 200,745 children (the control group C) were injected with a harmless saline solution (a placebo) and the other 201,229 children (the treatment group T) were injected with Salk's vaccine.

(a) What was the point of giving saline solution to the children who didn't get the vaccine? Explain briefly. *[10 points]*

(b) In experimental design, *double-blinding* is the process by which neither the subjects nor the people running the experiment know the treatment-control status of the subjects at the time the outcome of interest is measured for each subject. Would it have been possible to run this experiment in a double-blinded fashion? Would it have been a good idea to do so? Explain briefly. *[10 points]*

(c) The results of the trial were as follows: 33 of the 201,229 children who got the vaccine later developed paralytic polio, whereas 115 of the 200,745 saline children suffered this fate. Let $\hat{\theta}_T = \frac{33}{201229} \doteq 0.0001640$ and $\hat{\theta}_C = \frac{115}{200745} \doteq 0.0005729$ be the observed polio incidences in the T and C groups, respectively. Does the difference between these rates seem large to you in practical terms? Build a probability model for this situation, being explicit about all assumptions you make and why they're reasonable, and use your model to construct a 99.9% confidence interval for the population mean difference in rates of polio between the two groups. Sketch your confidence interval with $\left(\hat{\theta}_C - \hat{\theta}_T\right)$ as the center, locating the left and right endpoints, the center and the reference point of 0. Is the observed difference statistically significant at the 99.9% confidence level? What do you conclude about the effectiveness of the Salk vaccine? Explain briefly. *[70 points]*

(d) Your confidence interval sketch in (c) should have revealed that there was quite a bit of distance between the left endpoint and 0, which means that — in retrospect, after the experiment had finished — the designers of the trial had chosen T and C sample sizes that were quite a bit bigger than necessary. In the rest of this problem, let's roll the clock back to the period in which the trial was designed, and reconsider the sample size issue.

Let $n = (n_C + n_T)$ be the total sample size planned for the experiment, and for simplicity suppose that exactly $\frac{n}{2}$ children are randomized to each of the T and C groups. If the polio incidences turned out to precisely match the rates in the actual trial, what value of n would have been necessary to make the left edge of the 99.9% confidence interval be just barely positive? Show your work. (This method is one way to perform *sample size determination* at design time.) Do you think the designers of the Salk trial were stupid, or is there some other explanation for their retrospectively-unnecessarily-large sample sizes? Explain briefly. *[30 points]*