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

STAT 131: Take-Home Test 3, part 1 (required) *[270 total points]*

Due date: upload to `canvas.ucsc.edu` by **11.59pm Sun 14 Jun 2020**

Name: Christy Yuen

Note: Attended Professor Draper's Office Hours and got lots of help.

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]*

SOLUTION:

$Y_i = \theta + b + e_i$
θ and b are unobservable

$$\begin{aligned} e_i \sim IIDN(0, \sigma^2) & \\ E(Y_i) &= E(\theta + b + e_i) \\ &= \theta + b + E(e_i) \\ &= \theta + b + 0 \\ &= \theta + b \end{aligned}$$

$$
\begin{aligned}
V(Y_i) &= V(\theta + b + e_i) = V(e_i) = r^2 \\
Y_i &= \theta + b + e_i \\
Y_n &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(\theta + b + e_i) \\
&= (\frac{1}{n}\sum_{i=1}^{n}(\theta)) + (\frac{1}{n}\sum_{i=1}^{n}(b)) + (\frac{1}{n}\sum_{i=1}^{n}(e_i)) \\
&= \frac{n\theta}{n} + \frac{nb}{n} + (\frac{1}{n}\sum_{i=1}^{n}(e_i)) \\
&= \theta + b + e_n
\end{aligned}
$$

$$
\begin{aligned}
V(\bar{e}_n) &= V(\frac{1}{n}\sum_{i=1}^{n}(e_i)) \\
&= \frac{1}{n^2}V(\sum_{i=1}^{n}(e_i)) \\
&= IID\,\frac{1}{n^2}\sum_{i=1}^{n}V(e_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\end{aligned}
$$

$$
\begin{aligned}
V(\bar{e}_n) &= \frac{\sigma^2}{n} \\
E(\bar{Y}_n) &= E(\theta + b + \bar{e}_n) \\
&= \theta + b + E(\bar{e}_n) = \theta + b \\
E(e_n) &= E(\frac{1}{n}\sum_{i=1}^{n}(e_i)) \\
&= \frac{1}{n}E(\sum_{i=1}^{n}(e_i)) \\
&= \frac{1}{n}E(\sum_{i=1}^{n}0) = 0 \\
V(\bar{Y}_n) &= V(\theta + b + \bar{e}_n) \\
V(\bar{e}_n) &= \frac{\sigma^2}{n}
\end{aligned}
$$

$\bar{Y}_n$ has expected value $(\theta + b)$ and variance $\frac{\sigma^2}{n}$
$Y_i \sim IID$ mean $(\theta + b)$, variance $\sigma^2 < \infty$

$Y_n \sim IID$ mean $(\theta + b)$, variance $\frac{\sigma^2}{n}$

Weak Law of Large Numbers: $\bar{Y}_n \xrightarrow{\text{P}} \theta + b \neq \theta$ unless b=0 meaning the measuring is unbiased, which is not possible and not guaranteed.

$$
\begin{aligned}
RMSE\left(\bar{Y}_n\right) &= \sqrt{E(\bar{Y}_n - \theta)^2} \\
MSE(\bar{Y}_n) &= E[(Y_n - \theta)]^2 \\
&= E(\bar{Y}_n^2 - 2\theta\bar{Y} - n + \theta^2) \\
&= E(\bar{Y}_n^2) + E(-2\theta\bar{Y} - n) + E(\theta^2) \\
&= E(\bar{Y}_n^2) - 2\theta E(\bar{Y} - n) + \theta^2 \\
&= E(\bar{Y}_n^2) - 2\theta(\theta + b) + \theta^2 \\
fromearlier: & \\
V(\bar{Y}_n) &= E(\bar{Y}_n^2) - [E(\bar{Y}_n)]^2 \\
SoV(\bar{Y}_n) &= V(\bar{Y}_n) - [E(\bar{Y}_n)]^2 \\
&= \frac{\theta^2}{n} - [\theta + b]^2 \\
MSE(\bar{Y}_n) &= \frac{\sigma^2}{n} + (\theta + b)^2 - 2\theta^2 - 2\theta + \theta^2 \\
&= \frac{\sigma^2}{n} + b^2 \\
MSE(\bar{Y}_n) &= \frac{\sigma^2}{n} + b^2 \\
RMSE\left(\bar{Y}_n\right) &= \sqrt{E(\bar{Y}_n - \theta)^2}
\end{aligned}
$$

This is similar to Pythagoras's Expression and with more data, the smaller b will be.

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]*

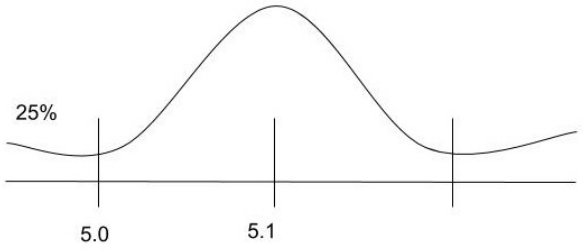
SOLUTION:

$SD(e_i) = \sigma = 0.15$

$b = 0$

$Y_i = 5.1 + 0 + e_i$

$e_i \, IID \, N(o, \sigma^2)$



PDF of $Y_i = 5.1 + 0 + e_i$

$$\frac{y-\mu}{\sigma} = \frac{5.0-5.1}{0.15} = \frac{-0.1}{0.15} = -0.67$$

$$\Phi(-0.67)$$

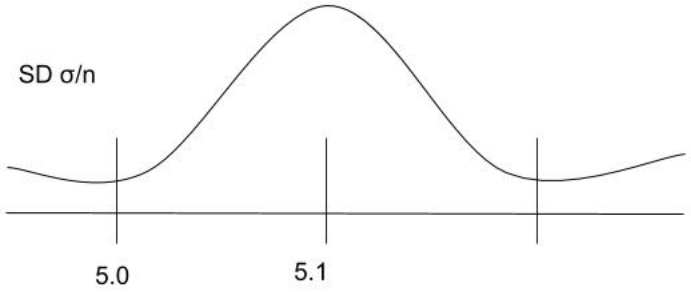
(c) You're not happy with the misclassification probability in (b), and you decide to remedy this by taking $n > 1$ independent water samples from the lake and basing your assessment on their mean pH value $\bar{Y}_n$. How large does n need to be to make the probability of {incorrectly concluding that 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]*

SOLUTION:

From WolframAlpha, P(we say threatened based on $\bar{Y}_n$, n = 1, truth = 5.0(not threatened))

= 25% (the misclassification error)

By making n(number of trials) bigger, the misclassification error rate will go down.



Looking for .0050, PDF of $\bar{Y}_n$, $n > 1$

$$\frac{y-\mu}{\frac{\sigma}{\sqrt{2}}} = \frac{5.0-5.1}{\frac{0.15}{\sqrt{n}}}$$

From WolframAlpha, I got -2.575

$$-2.575 = \frac{5.0-5.1}{\frac{0.15}{\sqrt{n}}}$$

$$n = 14.91$$

2. *[140 total points]* (medicine) Hypertension is a medical condition in which a person's blood

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.*

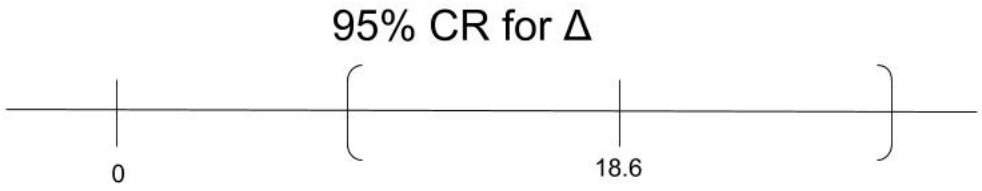
pressure is chronically elevated. (A reminder: blood pressure is measured with two numbers called *systolic* (higher) and *diastolic* (lower), in a deeply anachronistic scale called mmHg (millimeters of mercury); blood pressures are stored as data in the form "systolic over diastolic" [i.e., 115 over 75 or 115/75; 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."

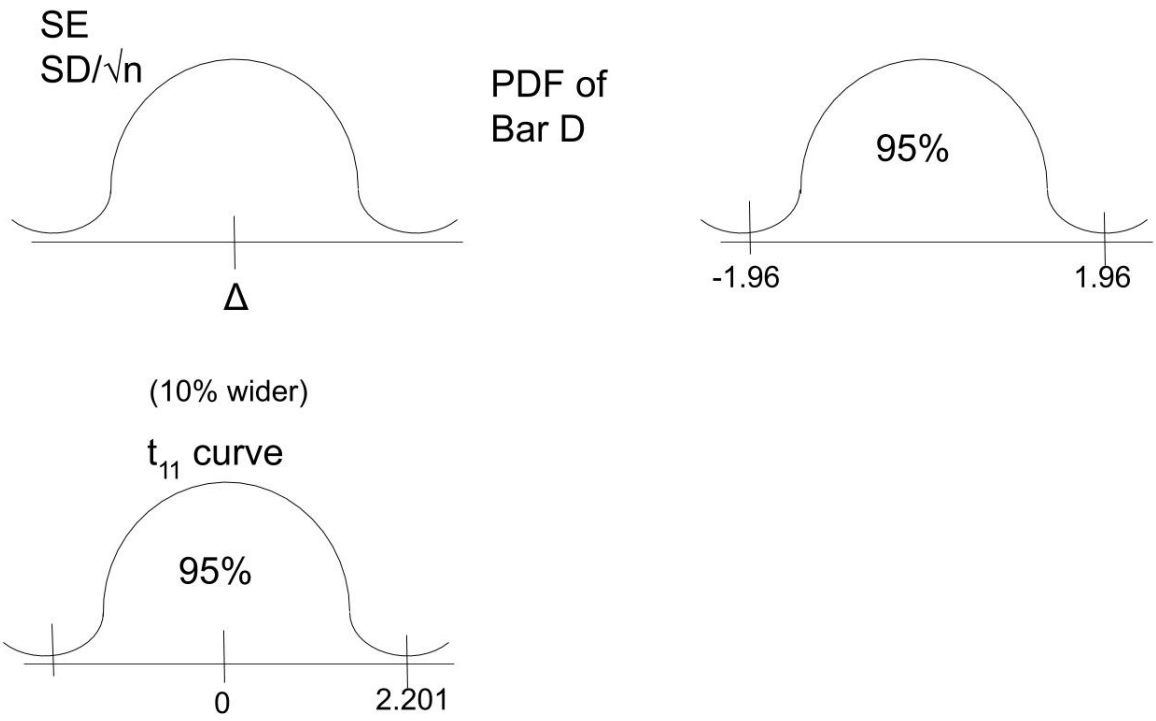
Captopril was developed in the mid 1970s; MacGregor et al. (1979, *British Medical Journal*) published the results of a clinical trial on its effects. Systolic blood pressures (in mmHg) were measured for $n = 12$ representatively-chosen hypertensive patients, before and after taking Captopril for a long 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]*

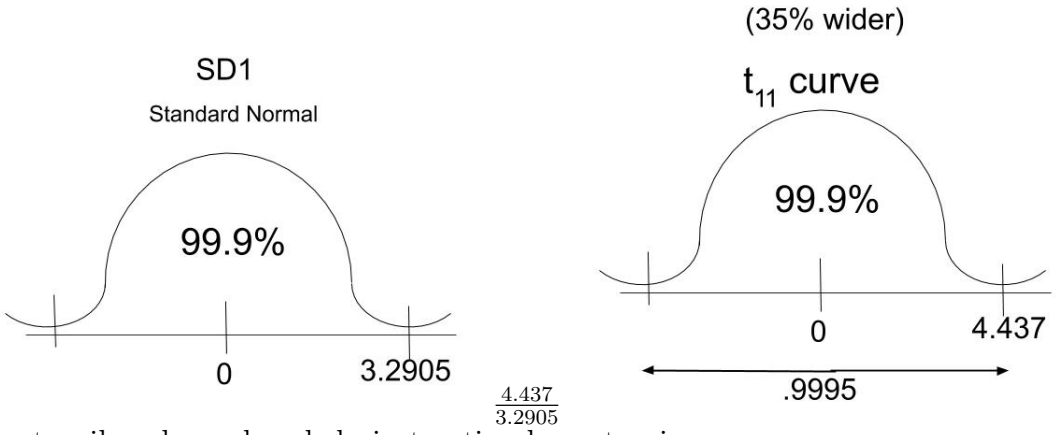
SOLUTION:



At 0, there is no effect on the average of Captopril on the population.
Since 0 is not in the 95% CI, the difference between $\hat{\delta} = 18.6$, and the no-average-effect of 0 is statistically significant implies it's hard to attribute to unlucky random sampling implies that it's possibly real.
Conclusion? Captopril works. The catch? This could be a false positive.
To decrease the chances of a false positive, we raise the confidence level from 95% to 99.9%. (Paraphrased from Professor Draper)



As n goes up, the t curves becomes more like the normal curve.



$\frac{4.437}{3.2905}$

Captopril perhaps does help in treating hypertension.

(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

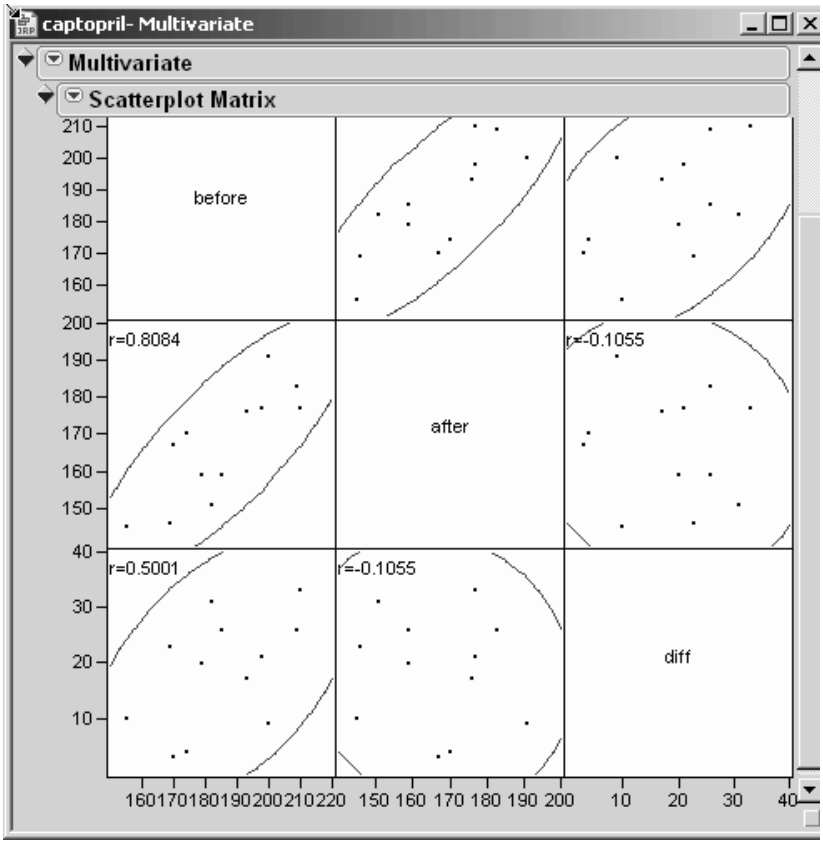


Figure 1: *Scatterplot matrix for the variables* `before`, `after`, *and* `diff`.

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 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]*

SOLUTION:

$$\begin{aligned} V_{CR}(\bar{D}_n) &= V_{CR}(\bar{B}_n - \bar{A}_n) \\ &\stackrel{I}{=} V_{CR}(\bar{B}_n) + (-1)^2 V_{CR}(\bar{A}_n) \end{aligned}$$

$$B_i \sim IID = \left\{ \begin{array}{l} E(B_i) = \mu \\ V(B_i) = \sigma_B^2 \end{array} \right.$$

$$A_i \sim IID = \left\{ \begin{array}{l} E(A_i) = \mu \\ V(A_i) = \sigma_A^2 \end{array} \right.$$

$$\begin{aligned} V_{CR}(\bar{B}_n) &= \frac{\sigma_B^2}{n} \\ \bar{A}_n &= \frac{1}{n}\sum_{i=1}^{n} A_i \\ V_{CR} &= \frac{\sigma_A^2}{n} \end{aligned}$$

So, $V_C R(\bar{D}_n) = \frac{\sigma_A^2+\sigma_B^2}{n}$

$$\begin{aligned} V_{RM} &= V_{RM}(\frac{1}{n}\sum_{i=1}^{n}[B_i - A_i]) \\ &= \frac{1}{n^2}V_{RM}[\sum_{i=1}^{n}[B_i - A_i]] \\ &= \frac{1}{n^2}[\sum_{i=1}^{n}V_{RM}[B_i - A_i]] \end{aligned}$$

$$\begin{aligned} V_{RM}(B_i - A_i) &= V_{RM}(B_i) + V_{RM}(A_i) + 2C_{RM}(B_i, -A_i) \\ &= V_{RM}(B_i) + (-1)^2 V_{RM}(A_i) - 2C_{RM}(B_i, -A_i) \end{aligned}$$

$$\frac{C_{RM}(B_i, -A_i)}{SD(A_i) * SD(B_i)} = \rho_{AB}$$

$$\begin{aligned} V_{RM} &= \frac{1}{n^2}\sum_{i=1}^{n}V_{RM}(B_i - A_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}[\frac{\sigma_A^2}{V(A_i)} + \frac{\sigma_B^2}{V(B_i)}] - 2\rho_{AB}\sigma_A\sigma_B \\ &= \frac{1}{n^2}\sum_{i=1}^{n}(\sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B) \\ &= \frac{\sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B}{n} \end{aligned} \tag{6}$$

$2\rho_{AB} = 0.8084$

(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]*

SOLUTION:
The drug doesn't seem constant with the 12 patients as there is a tendency for the higher the patient's before number, the better the improvement. Everyone improved under Captopril but the difference numbers range from +3 to +33. This is a really big difference as everyone had different improvements. Patient 4's before number is unusual as it's the 3rd smallest before number and had a small improvement. While patient 12 has a large before number and a large improvement. It's not always the case but that