Philosophy of Statistics: Homework 5

due on Gradescope by 11am on Thursday June 3

Guidelines. Some questions ask you to justify your answers. For these questions, credit will be based on how well you justify your answers, not on whether your answers are correct. (There's often no consensus on the correct answers, even among statisticians.) However, that doesn't mean that anything goes: some answers will be hard to justify well. I give suggested word counts but these are just ballpark numbers. Don't sweat them too much. Collaboration is encouraged, but make sure to write up your answers by yourself and list your collaborators.

Problem 1 (15 points). I hard-boil eggs by putting them in a pot of cold water, bringing it to the boil, letting it simmer for seven minutes, draining the eggs, and peeling them. Some of the eggs are easy to peel: the shell comes off quickly and cleanly. But others are hard to peel: it takes ages and chunks of egg come off with the shell despite my best efforts.

A friend suggests an alternative method: bring the water to the boil first, then add the eggs and simmer for ten minutes. They claim that this method tends to make the eggs easier to peel.

To test their claim I plan to buy two cartons of eggs. I'll hard-boil the eggs from one carton using my method and from the other carton using their method. If more eggs from the second batch are easy to peel, I'll conclude that my friend is right.

It's possible that this procedure leads me to the wrong conclusion. For example, suppose the second carton's eggs are fresher than the first's, and fresher eggs are easier to peel. Then, even if my friend is wrong, I will end up concluding that they're right.

Describe five more scenarios in which the procedure will lead me to the wrong conclusion.

Problem 2 (15 points). On the standard view, we should allocate experimental subjects to treatment and control groups at random. But what does randomization achieve? Some suggest that, by randomizing, the treatment and control groups will tend to have a similar distribution of prognostic factors. For example, Tanur et al. (1989: 10) say that randomization "usually will do a good job of evening out all the variables—those we didn't recognize in advance as well as those we did recognize."

In class, we examined this claim using the Legos example. First, explain in detail how the Legos example works. Second, discuss whether the Legos example refutes the claim. You might like to consider: (a) whether the Legos example is an appropriate analogy; (b) whether there are alternative interpretations of the claim, interpretations not undermined by the Legos analogy. (300 words)

Problem 3 (15 points). Suppose your friend claims they can taste the difference between Coke and Diet Coke. To test their claim, you blindfold them and have them taste ten cans—five of Coke and five of Diet Coke—and see how many they get right. If they get, say, nine or ten right, that's pretty good evidence they were telling the truth.

OK, but how should you order the cans? A few options:

(a) c, c, c, c, c, dc, dc, dc, dc, dc

(b) c, dc, c, dc, c, dc, c, dc, c, dc

(c) randomly choose one of the 252 possible orderings

(d) choose an order you're confident your friend won't be able to predict

Discuss the pros and cons of these options (300 words). Things you might like to think about: (i) What if your friend guesses correctly not because they taste the difference but because they predicted the order? (ii) What if the randomly chosen order in (c) turns out just to be the order in (a)? (iii) What if you don't know your friend as well as you think you do?

Problem 4 (10 points) The following questions are about two of the readings: Bartlett et al.'s *Extracorporeal Circulation...*, and Chapter 1 of Kulkarni and Harman's *Statistical Learning Theory.*

1. What were the results of the phase 1 trial of ECMO?
2. Describe the "randomized play-the-winner" technique used in the study.
3. Summarize the reasons given by the authors for using the randomized play-the-winner technique.
4. The families of infants in the study were asked to consent to the selected treatment only if what?
5. What were the study's results?
6. In statistical learning theory, what is a feature vector?
7. How do Kulkarni and Harman define inductive learning?
8. What does the term "supervised" in "supervised learning" refer to?

Problem 5 (25 points). Writing good homework questions is tricky but instructive: by writing them you improve your understanding of the topic, or so I've found. Give it a go!

Your question can be about any of the topics we've covered. It can take any form you like. For example, it can test comprehension of one of the readings, or require some calculation, or ask for discussion of some conceptual issue, or so on. It shouldn't just be a slight variation on a previous homework question. It should be worth between 10 and 20 points. Include a model answer to your question.

Problem 6 (20 points). We've covered a variety of topics: simplicity, algorithmic fairness, merely statistical evidence, correlation v. causation, randomized controlled trials, and statistical learning theory. Quite a list! In the final class on Thursday, we'll review the key ideas together and try to get some perspective on the course. To prepare for that, pick *two* of these topics and review the material. Then:

(a) summarize the key questions, ideas, proposed solutions

(b) describe anything you're confused about

(c) if your views about the topic changed, describe how

Be prepared to discuss in the final class. (300 words for each topic.)

That's it for the homeworks. You made it. Well done!