Philosophy of Statistics: Homework 8

due on Gradescope by 11am on Friday March 5

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 (20 points). Imagine Jim and Pam are Bayesians conducting an experiment together. We can distinguish a few ways in which, after seeing the result of the experiment, Jim and Pam might disagree:

1. They might have different posteriors in the hypotheses.
2. They might disagree about the strength of evidence of the result for a hypothesis.
3. They might disagree about whether the result is evidence for or against a hypothesis.

Give an example for each case. So, for each case, your answer should consist of a table showing the possible outcomes, their likelihoods under the hypotheses, Jim's priors, Pam's priors, and a calculation showing that your example meets the relevant condition. (To measure the strength of evidence, use the typical measure: the difference between posterior and prior. Don't use any example discussed in class.)

Problem 2 (20 points). In a previous homework, we considered the relationship between the *reliability* of a procedure and how *confident* you should be in the procedure's output. How? By imagining a machine for evaluating statements: you feed in the statement, the machine churns away for a while, and then it prints out 'True' or 'False'. Even supposing the machine is reliable, it doesn't follow that you should be confident its answer in a particular case is correct. (To take an extreme case: if you fed in the statement '$3 + 4 = 8$' and the machine printed out 'True', which is perfectly consistent with the machine's being reliable, you would be certain that its answer in this case was incorrect.)

Now let's consider the relationship between the *reliability* of a procedure and the *evidential import* of the procedure's output. How? By again imagining a machine for evaluating statements. But this machine works slightly differently: you feed in the statement, the machine rolls a 100-sided die, and then it prints out 'True' or 'False' along with the result of the die roll. Now, if the die lands on 1–99, it prints out the correct answer ('True' if the statement you fed in is true, 'False' if it's false); but if the die lands on 100, the machine prints out 'True' or 'False' at random. So the machine is reliable: with high probability, it prints out the correct answer. But it does not follow that its printing 'True' is always evidence that the statement you fed in is true or that its printing 'False' is always evidence that the statement you fed in is false!

First, explain why not, with examples. Second, show that if we modified how the machine worked there are cases where its printing out 'True' is evidence that the statement you fed in is in fact *false*,

and that its printing out 'False' is evidence that the statement you fed in is in fact *true*.

In short: in the last homework we showed that *reliability* doesn't always yield *confidence*; here we have shown that *reliability* doesn't always yield *evidence* either.

Problem 3 (30 points) What's the point of thinking about toy examples of reliable statement-evaluating machines, as we've been doing? The point is that these toy examples help us evaluate the long-run defense of frequentism.

One kind of long-run defense says roughly: "My procedure for evaluating hypotheses is reliable, so you should be confident that the procedure's output in this case is correct." Another kind of long-run defense says roughly: "My procedure for evaluating hypotheses is reliable, so the procedure's output is evidence about the hypotheses."

First, explain in more detail what the long-run defenses of frequentism say. (Be as specific as you can. You should talk about both significance tests and Neyman-Pearson tests.) Second, explain how someone might attempt to use the examples of the reliable statement-evaluating machines to refute the long-run defenses. Third, discuss whether that attempt succeeds. (300 words.)

Problem 4 (15 points). A Bayesian says that an outcome x is evidence for a hypothesis h just if the posterior is greater than the prior. In symbols: $P(h \mid x) > P(h)$. OK, but how should we measure the *strength* of the evidence? In a previous homework, we considered two possible measures:

> Difference measure: the strength of evidence of x for h is measured by the *difference* between posterior and prior, i.e. $P(h \mid x) - P(h)$.
>
> Ratio measure: the strength of evidence of x for h is measured by the *ratio* of posterior and prior, i.e. $P(h \mid x)/P(h)$

And you showed that these measures disagree. But yet other measures are possible. For example:

> Likelihood ratio measure: the strength of evidence of x for h is measured by the likelihood ratio for h over $\neg h$, i.e. $P(x \mid h)/P(x \mid \neg h)$.

Show that the difference measure and the likelihood ratio measure disagree too. That is: write down an experiment where x_1 is evidence for h and x_2 is evidence i, but according to one of the measures, the strength of evidence of x_1 for h is greater than the strength of evidence of x_2 for i, and according to the other measure, the strength of evidence of x_1 for h is less than the strength of evidence of x_2 for i.

Problem 5 (15 points). In previous homeworks, we've come up with examples where different approaches to evaluating statistical hypotheses give contradictory conclusions. For example, in Problem 5 of Homework 6, you showed that there are cases where the significance tester says the outcome is evidence *against* a hypothesis but the Bayesian says the outcome is evidence *for* that hypothesis. Take one of the examples of disagreement you came up with: any you like, or a fresh example if you prefer. Write down the example again. Which of the approaches, if either, gets it right? (200 words.)