Philosophy of Statistics: Homework 7

due on Gradescope by 11am on Friday February 26

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). Imagine 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'. The machine is very reliable: if you feed in a true statement, then with high probability it prints 'True'; if you feed in a false statement, then with high probability it prints 'False'. But it does not follow that you should be very confident, of any particular answer printed by the machine, that the answer is correct! Give an example where you would be certain the answer was correct, an example where you would be very confident the answer was correct, an example where you would be very confident the answer was incorrect, and an example where you would be certain the answer was incorrect.

Problem 2 (20 points). A stopping rule tells you what to *do* in an experiment. For example, in an experiment to test whether a coin is fair one stopping rule is "flip the coin 20 times" and another stopping rule is "flip the coin until you get a total of 6 tails". There are many other possible stopping rules.

Now, imagine a basketball coach wishes to test whether a player's free throw percentage is 80%, as the player claims. So the coach has the player take some free throws. The results are: make, miss, miss, make, make, miss, make, make, make, miss, make, miss, make, make, make. Give four examples of stopping rules which the coach might be using. (Don't worry about whether your stopping rules are weird or not. Any four stopping rules consistent with the results will do.)

Problem 3 (15 points). In class, we discussed how significance tests are *sensitive to the stopping rule*. Explain what this means and illustrate with an example. (You can use the example we discussed in class if you like, or come up with your own.) Be as precise as you can.

Problem 4 (25 points). Howson and Urbach (2006: 158–9) say: "The fact that significance tests and, indeed, all classical inference models [are sensitive to the stopping rule] is a decisive objection to the whole approach." Royall (1997: 24) agrees, saying that sensitivity to the stopping rule is a "structural flaw" and means that frequentist methods are "invalid".

Let h be some hypothesis and x be the outcome of an experiment in a significance test of h. We might try to formulate the objection as an argument, like this:

(P1) The strength of evidence of x against h doesn't depend on the stopping rule.

(P2) The p-value of x does depend on the stopping rule.

(C) So, the p-value doesn't measure the strength of evidence of x against h.

Evaluate this argument. Be as specific as you can. If a premise is false, which premise and why? If the argument is invalid, how does it go wrong and how might it be improved? If the conclusion is true, can anything be salvaged from significance testing? (300 words.)

Problem 5 (25 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 class we measured it like this:

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

But other measures are possible. For example:

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

These two measures are quite different. For example, there are experiments where each of two of the possible outcomes, x_1 and x_2, are evidence for each of two hypotheses, h and 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. Given an example of such an experiment and show your calculation of the strengths of evidence.