Philosophy of Statistics: Homework 6

due on Gradescope by 11am on Wednesday February 17

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). How do you tell whether a method of evaluating statistical hypotheses is any good? A sensible way to start, I suggested, is to try out the method in various cases and see if it delivers the correct conclusions. Just as to tell if an omelet recipe is any good, you can follow the recipe and see how the omelet tastes, so too to tell if a method of evaluating statistical hypotheses is any good, you can follow the method and see how the conclusions look.

Allen asked an excellent question on the forum about this "omelet" approach:

> To find out if an omelet recipe is good, we can just try it out and see how the omelet tastes. Supposedly similarly, to find out if a method of evaluating statistical hypothesis is good, we can try it out and see if the results are good. But how do we know if the results are good? It seems that there is no way to "taste" the result. It seems that we can only compare it to results from other methods, but that would just be like comparing omelets from other recipes without being able to taste them.

Is Allen right that the "omelet" approach is hopeless? You might find it helpful to look at my reply to Allen's post on the forum. But my reply only scratches the surface, so even if you agree with it (which you may not!) try to develop it further. (300 words.)

Problem 2 (15 points). In Class 10, we saw examples of significance tests where the distributions over the outcome spaces were very different but all the significance tests resulted in the same p-value. This question asks you to develop your own such examples.

Give an example such that the p-value is .05 and...

1. there are only two possible outcomes
2. the actual outcome has probability .05 and no outcome has probability above .06
3. there are eight possible outcomes and the actual outcome is the second most likely
4. the actual outcome has probability .000001

Each example should specify: a probability distribution, the actual outcome, and a calculation of the p-value.

Problem 3 (15 points). The p-value of an outcome is the probability, assuming the null hypothesis is true, of getting an outcome as extreme as the actual outcome. But what does *as extreme* mean? In class, we defined it like this:

1. o_1 is as extreme as o_2 just if $\mathrm{P}_h(X = o_1) \leq \mathrm{P}_h(X = o_2)$

Scott suggested an alternative definition:

2. o_1 is as extreme as o_2 just if $\mathrm{P}_h(X = o_1)$ is at least as far away from the average of all the probabilities as is $\mathrm{P}_h(X = o_2)$.

First, give an example where the two definitions yield different p-values. Second, give an example where the two definitions yield the same p-value. (Each example should specify: a probability distribution, the actual outcome, and a calculation of the two p-values. Avoid trivial examples, e.g. examples where there's only one outcome.) Third, discuss whether the second definition is an improvement on the first. (300 words.)

Problem 4 (25 points). Let's compare and contrast significance tests and Neyman-Pearson tests.

Consider the following cases:

1. The outcome has a lower p-value under i than under h, yet it would lead us to reject h in favor of i in a Neyman-Pearson test using rejection region R_1.
2. The outcome is the least likely outcome under i, yet it would lead us to reject h in favor of i in a Neyman-Pearson test using rejection region R_2.
3. The probability of mistaken rejection in a significance test of h, with significance level .05, is equal to the size of R_1, yet the set of outcomes which would lead us to reject h in the significance test and the set of outcomes which would lead us to reject h in the Neyman-Pearson test have no outcomes in common.

In each case, either give an example or show that none exists. In your examples the rejection regions should not be the trivial regions: the empty region or the set of all outcomes. As usual, you should assume that the Neyman-Pearson tester is testing h against i, so h is the null and i is the alternative.

The upshot: the significance tester and the Neyman-Pearson tester disagree, and how!

Problem 5 (25 points). Let's compare and contrast significance tests and Bayesianism. Can you come up with examples where...

1. the outcome's p-value is .05, for the significance tester, but the Bayesian says the outcome is evidence *for* that hypothesis?
2. the outcome's p-value is 1, for the significance tester, but the Bayesian says the outcome is evidence *against* that hypothesis?
3. the significance tester says that Outcome 1 is stronger evidence against the hypothesis than Outcome 2, but the Bayesian says the reverse?
4. the outcome which the significance tester takes to be the strongest evidence *against* the hypothesis is the outcome which the Bayesian takes to be the strongest evidence *for* that hypothesis?

Each example should specify: a table showing the Bayesian's priors and likelihoods, which of the hypotheses is the significance tester's null hypothesis, and a calculation or argument to show that your example satisfies the conditions.

The upshot: the significance tester and the Bayesian disagree, and how!