Philosophy of Statistics: Homework 3

due on Gradescope by 11am on Wednesday January 27

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). Compare two forms of argument:

> *Modus Tollens (MT).* If H is true then r won't happen. In fact, r happened. That is conclusive evidence against H.
>
> *Probabilistic Modus Tollens (PMT).* If H is true, then r probably won't happen. In fact, r happened. That is good evidence against H.

MT is a sound form of argument: in any instance of MT, if the premises are true so is the conclusion. PMT is similar to MT, so you might be tempted to think it's sound too. But it's not. Come up with three instances of PMT where the premises are true but the conclusion is false. (To show that the conclusion of an instance of PMT is false, don't do any significance tests. Just use your own judgment.)

Problem 2 (15 points). Here is a probability distribution in an experiment, assuming that the null hypothesis is true:

1	2	3	4	5	6	7	8
.001	.340	.009	.010	.527	.040	.051	.022

Suppose we do a significance test (v.2) using a significance level of .05 and sticking with the definition of *as extreme* from class.

First, for which outcomes would we end up believing the null hypothesis is false in our significance test? Second, the probability of *mistaken rejection* in a significance test is the probability, assuming the null hypothesis is true, of ending up believing that the null hypothesis is false. What is the probability of mistaken rejection in our significance test?

Problem 3 (20 points). In the previous question, the probability of mistaken rejection in the significance test is less than or equal to the significance level. (Check your working if not.) You might wonder whether that's always the case. So either show, by giving an example, that the probability of mistaken rejection in a significance test can be greater than the significance level, or show that no such example exists.

Problem 4 (15 points). In any significance test, the first step is to design an experiment: what to *do* and what to *count*. To test whether a coin is fair (i.e. equally likely to land heads as tails when flipped), there's a pretty natural experimental design: what to do is flip the coin, say, 100 times and what to count is the number of heads. All nice and straightforward. But sometimes it's not so clear what to do or what to count. For example, think about a die instead of a coin: it's not so clear what to do or what to count in order to test whether a die is fair (i.e. equally likely to land on each face when rolled).

Let's just stipulate that we'll roll the die, say, 600 times. That settles the question of what to *do*. But it doesn't settle the question of what to *count*. We could count the number of 6's, in which case the possible outcomes of the experiment would be 0, 1, 2, ..., 600. Or we could count the number of 2's, in which case the possible outcomes of the experiment would again be 0, 1, 2, ..., 600. Or we could count the sum of the rolls, in which case the possible outcomes of the experiment would be 600, 601, 602, ..., 3600. Or we could count the average of the 600 rolls, in which case the possible outcomes of the experiment would include, for example, 1.5, 3.511, 5, and many more. There are lots of things we could count.

For any particular sequence of 600 rolls of the die (e.g. 1, 4, 3, 3, 5, 6, 6, ...), the p-value is *sensitive* to what we count: you might get a p-value of, say, .04 if you count one way and, say, .13 if you count another way.

So the question is: *What should we count?* Maybe you endorse one of the examples above. Or maybe you have your own example. Or maybe you think it doesn't matter what we count. Or maybe you think there's no basis for choosing what to count, and that's a problem for significance testing. Or maybe you think something else. In any case, justify your answer. (300 words)

Problem 5 (20 points). There are a lot of new concepts in Neyman-Pearson testing: *rejection region*, *size*, *power*, *domination*, and *likelihood ratio tests*. This question is designed to help you get comfortable with these concepts.

Here are two probability distributions for an experiment:

	1	2	3	4
h	.1	.3	.05	.55
i	.05	.15	.65	.15

The experiment has 4 possible outcomes: 1, 2, 3, and 4. The table shows the probability of each outcome according to h, the null hypothesis, and according to i, the alternative hypothesis.

(a) How many possible rejection regions are there?

(b) What is the size and power of $\{1, 3\}$?

(c) Which rejection regions are undominated?

(d) Which rejection regions are likelihood ratio tests?

(e) Among undominated rejection regions with size at most 0.2, which has the highest power?

Problem 6 (15 points). In Neyman-Pearson tests, we *decide between* two hypotheses. But is it sensible to decide between hypotheses like this? Perhaps we could just come to some judgment about their relative plausibility instead. A defender of Neyman-Pearson testing might reply:

> We have no option but to decide between hypotheses. In industrial quality control, for example, you must either ship the product or not ship it. Or in drug testing, you must either approve the drug or not approve it. Or in the Tulips example from class, you must either label the box of bulbs as 75% red-flowering, or not so label it. No intermediate response is possible! So the fact that Neyman-Pearson tests involve deciding between hypotheses is not a problem. Far from it: there is simply no other option.

Evaluate this reply. (300 words)