DS 100: Principles and Techniques of Data Science **Date: April 20, 2018**

Discussion #11

Name:

Hypothesis Testing

1. In lecture, we saw an example of permutation inference in Boring, Ottoboni, and Stark's (2016) reexamination of Student Evaluation of Teaching (SET) data. In the experiment, 47 students were randomly assigned to one of four sections. In two of the sections, the teaching assistants were introduced using their actual names. In the other two sections, the assistants switched names. Students never met the teaching assistants face-to-face. Instead, they interacted with the students mainly via an online forum. Homework returns were coordinated so that all students received feedback all at the same time. The authors wanted to investigate if gender perception has any effect on SETs.

   (a) What is the model?

   Solution: Each TA has two possible ratings from each student—one for each perceived gender. Each student had an equal chance of being assigned to any one of the (gender, perceived gender) pairs. The students evaluate their TAs independently of one another.

   (b) What is the null hypothesis?

   Solution: Perceived gender has no effect on SETs. In the model, this means that each TA really only has one possible rating from each student.

   (c) What is the test statistic?

   Solution: The difference in means of perceived male and perceived female.

   (d) How did the authors use permutation to compute the sampling distribution under the null hypothesis?

Solution: They permuted the perceived gender labels for students under the same TA and calculated the test statistic.

(e) How is this permutation justified?

Solution: Under the null model, each student would have given their TA the same promptness rating regardless of perceived gender. Simple random assignment then implies that for a given TA, all their ratings had an equal chance of showing up under perceived male or perceived female.

(f) The TAs objectively returned the assignments at the same time. Therefore, assigned TA should also have no effect on promptness. Why, then, do the authors not permute across TA assignments?

Solution: Students might have given TA1 and TA2 different ratings based off other aspects of their interactions (spillover effects).

2. Suppose we roll a die 10000 times. The first 5000 rolls are done while wearing a fedora, and the latter 5000 rolls are done while wearing a 10-gallon hat. The type of hat does not affect the die. However, we would expect that the null hypothesis—exactly as many even numbers will be rolled while wearing a fedora as while wearing the 10-gallon hat—would not likely be true. Stated differently, random fluctuations in the proportion of even rolls while wearing the different hats imply that the null hypothesis will nearly always be false.

   Comment on the validity of the claim above.

   Solution: There's something amiss here. The null hypothesis should be a statement about a model. For example, all rolls of the dice can be modeled as independent coin flips with heads representing even rolls and tails representing odd rolls. The flips have probabilities p_{fedora} and $p_{ten-gallon}$ of showing heads while wearing the fedora and ten-gallon hat, respectively. The null hypothesis is that $p_{fedora} = p_{ten-gallon}$. Here the probabilities are fixed unknown numbers—they are not proportions taken from the data. We can use the observed frequencies as evidence for or against the null model.

3. The Graduate Division at UC Berkeley compares admission rates for men and women. For one year and one graduate program, this is summarized in the following table:

	Admitted	Rejected	
Men	509	316	825
Women	89	19	108
	598	335	933

Either argue against the sensibility of the following question or describe a simulation or resampling method that can be used to address it: Is the difference between admission rates for men and women statistically significant?

Solution: The table and question are suggestive of a simulation where we fix the number of men and women applying to the program and randomly draw for admission or rejection. However, before we set up that machinery, we should pause and think about the statistical question we would be answering.

We can't identify a pool of potential applicants. And even if we could, the applicants are not drawn from this shortlist by any probability method. This leads to the fixed margins. Therefore the variability in the simulation should correspond to the admission process, but we already know that departments don't admit candidates by drawing names from a hat, so eliminating chance as an explanation for the difference is a vacuous task.

4. Saccharin is used as an artificial low-calorie sweetener in diet soft drinks. There is some concern that it may cause cancer. Investigators performed an experiment on rats. In the treatment group, the animals got 2% of their daily food intake in the form of saccharin. The treatment group had a higher rate of bladder cancer than the control group, and the difference was highly significant (p=0.01). The investigators concluded that saccharin probably causes cancer in humans. Is this a good way to interpret the p-value? If not, then what, if anything, can the p-value say about this experiment?

Solution: No. The p-value tells you that the higher rate of cancer is unlikely to be a fluke in the random assignment of animals to treatment or control. The p-value does not help you extrapolate from high doses in rats to low doses in humans. Making that claim would require a domain-based argument. Clinical research is typically conducted in successive stages, see en.wikipedia.org/wiki/Clinical$_{t}rialPhasesforadescription$.