Data 8 Spring 2016 hw08, Due: 5pm Thursday, April 21

Please write your answers in the space provided. You can write on a printed copy or fill in the blanks with a PDF editor such as Acrobat Reader or Apple Preview. (*Beware: Some people have found that editing a PDF in a browser doesn't work.*) When you're done, upload a scanned copy to Gradescope (`gradescope.com`). You are welcome to use `ds8.berkeley.edu` to try out Python expressions. Directly sharing answers is not okay, but discussing problems with course staff or students is encouraged.

This assignment is due 5pm Thursday, April 21. You will receive an early submission bonus point if you turn it in by 5pm Wednesday, April 20.

Problem 1 *Another Way to Compute Confidence Intervals*

A Data 8 student suggests a new way to form confidence intervals for estimating a parameter from a statistic, using two concepts from earlier in the semester:

- Roughly 95% of the values in a normal distribution lie within two standard deviations of the mean.
- The variability of a statistic computed for resampled samples approximates the variability of that statistic computed for samples from the population.

With these ideas in mind, the student proposes the following method:

1. Take a random sample of size N from the population and compute the statistic T on the sample.
2. Resample N elements at random with replacement from this sample many times and compute the same statistic T on each resampled sample.
3. Compute the standard deviation S_{boot} of the resample statisics (boot stands for "bootstrap").
4. Form a confidence interval from $T_{\text{sample}} - 2S_{\text{boot}}$ to $T_{\text{sample}} + 2S_{\text{boot}}$

(a) List **all** of the condition(s) below that must hold for this method to compute a 95% confidence interval:

(i) the sample includes a large proportion of the population
(ii) the population distribution is a normal (bell-shaped) distribution
(iii) the sampling distribution of T is a normal (bell-shaped) distribution
(iv) the values are in standard units

1(a):

(b) List all of the estimation problems below for which this method would apply:

(i) Estimating the mean of any population distribution, using the mean of a sample
(ii) Estimating the mean of a normal population distribution, using the mean of a sample
(iii) Estimating the max of any population distribution, using the max of a sample
(iv) Estimating the max K of a uniform population distribution from 1 to K, using twice the mean of a sample

1(b):

(c) How could we modify this method to compute a 99.7% confidence interval?
1(c):

Answer:

(a) (iii)

(b) (i), (ii), (iv) OR just (i) and (ii); (iv) technically computes an estimator for $K+1$ instead of K.

(c) Since 99.7% of the values in a normal distribution lie within 3 SD of the mean, we could replace the 2s with 3s.

Problem 2 *Formulating Hypothesis Tests*

State a null hypothesis for each of the following test scenarios.

(a) **Tires:** Turtles Tires claims that their tires last for at least 50,000 miles. We believe that the average tire lifetime is less that 50,000 miles and we'd like to test this belief.
2(a):

(b) **Body Temperature:** A "normal" human body temperature has long been considered to be 98.6 degrees Fahrenheit. In a group of 100 people sampled at random, the average temperature was found to be 98.4. We'd like to test whether or not this difference is statistically significant.
2(b):

(c) **Knee Surgery Recovery:** A researcher is studying the effects of an innovative exercise program on knee surgery patients. There is a chance that the therapy will improve recovery time, but there's also a chance it will make it worse. A standard recovery time for knee surgery patients (without the exercise program) is 8.2 weeks.
2(c):

Answer: A null hypothesis is a statement about the opposite of what you're testing for.

(a) average tire lifetime is greater than or equal to 50,000 miles.

(b) The average temperature of the group is 98.6 degrees.

(c) The average recovery time for patients receiving the innovative exercise program is 8.2 weeks.

Problem 3 *Evaluating Hypothesis Tests*

Indicate whether each of the following statements is True or False.

1. True / False: A p-value is the probability that the null hypothesis is true.
   3(a):
2. True / False: A very small p-value gives evidence in support of the null hypothesis.
   3(b):
3. True / False: For a hypothesis test we find a p-value of .02, and use a significance level of 0.05. In this case we would reject the null hypothesis.
   3(c):
4. True / False: For a hypothesis test we find a p-value of .02, and use a significance level of 0.01. In this case we would reject the null hypothesis.
   3(d):
5. True / False: A p-value is always between 0 and 1 (inclusive).
   3(e):

Answer:

1. False: A p-value is a conditional probability – given the null hypothesis is true, it's the probability of getting a test statistic as extreme or more extreme than the calculated test statistic.

2. False: A very small p-value provides evidence *against* the null and in support of the alternative hypothesis.
3. True: If p-value $< \alpha$, we reject the null.
4. False: In this case we would fail to reject the null hypothesis.
5. True. A p-value is a probability, so it will always be between 0 and 1.

Problem 4 Dubious Dice

Students in a Data Science class are testing whether a die is fair or not. That is, they are testing whether each face of the die appears with chance 1/6 on each roll, regardless of the results of other rolls. The die is rolled n times. The counts of each face are stored in an array `c` with six elements. For example, `c.item(2)` is the number of times that the outcome was 3. `sum(c)` is n.

Write a Python expression that computes the total variation distance between the empirical distribution of outcomes and the probability distribution of a fair die.
4:

Answer: `sum(abs(c/n - 1/6)) * 0.5`