DS 100: Principles and Techniques of Data Science **Date: April 13, 2018**

Discussion #10

Name:

Hypothesis Testing

1. Define these terms below as they relate to hypothesis testing.
   (a) Data Generation Model:

   (b) Null Hypothesis:

   (c) Test Statistic:

   (d) Sampling distribution

   (e) p-value:

2. State whether each statement below is True or False. Provide an explanation.
   (a) p-values can indicate how incompatible the data are with a specified statistical model.

   (b) p-values measure the probability that the null hypothesis is true.

   (c) If our p-value is small, we have proven that the null model is false.

   (d) The p-value is the probability of the null hypothesis given the data.

   (e) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Bootstrap

We take an i.i.d. random sample of size 9 from a population. We write all the values on pieces of paper and stick them in a box:

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{2}\ \boxed{3}\ \boxed{3}\ \boxed{3}\ \boxed{4}\ \boxed{4}\ \boxed{5}}$$

The numbers in the box have the following summary statistics:

Statistic	Sum	Sum of Squares	Mean	Median
Value	27	93	3	3

3. For each of the following, answer the following questions: Is this value calculable from the information given? If so, either calculate it by hand or describe how you would calculate this value. If not, then suggest an estimate for the quantity. All draws are with replacement.
   (a) The expected value of a single draw from the box.

   (b) The expected value of the average of nine draws from this box

   (c) The exact variance of the tickets in the box

   (d) The exact variance of a single draw from the box

   (e) The exact variance of the average of nine draws from the box

   (f) The exact variance of the average of nine draws from the population

4. Let's say we forgot the analytic solution for finding the variance of the average of nine draws with replacement from the population. Describe a bootstrap procedure to estimate the variance.

5. What are the sources of error in the bootstrap procedure?

6. Which of the following could be valid bootstrap resamples? Provide reasons for the ones that are not.
   (a) 1, 2, 2, 3, 3, 4, 4, 5, 6
   (b) 1, 2, 2, 2, 3, 3, 3, 4, 4, 5
   (c) 1, 1, 1, 1, 1, 1, 1, 1, 1
   (d) 2, 2, 3, 3, 3, 4, 4, 4
   (e) 1, 2, 3, 3, 3, 4, 4, 5, 5

7. What are some assumptions we are making when performing the bootstrap?

8. You generate 10 bootstrap resamples (you would normally take many more). They are sorted and printed below:

```
[1, 2, 2, 2, 4, 4, 4, 4, 5] [1, 2, 3, 3, 3, 3, 3, 4, 4]
[1, 2, 2, 3, 3, 3, 4, 4, 5] [1, 1, 2, 3, 4, 4, 4, 5, 5]
[2, 3, 3, 3, 4, 4, 4, 5, 5] [2, 3, 3, 3, 3, 3, 3, 4, 4]
[1, 1, 1, 1, 2, 2, 3, 4, 5] [2, 2, 3, 4, 4, 4, 4, 4, 5]
[1, 2, 2, 3, 3, 3, 4, 4, 4] [1, 2, 2, 2, 3, 3, 3, 4, 5]
```

   Construct a 60% confidence interval for the population 40^{th} percentile of the population.

9. Which of the following statements are valid claims? Provide revisions for the others.
   (a) There's a 60% chance that the confidence interval in question 6 covers the true population 40th percentile.
   (b) If we were to repeat our sampling procedure and bootstrap confidence interval estimation many times on the population, then in the limit of infinite samples, at least 40% of those 60% confidence intervals will cover the 40th percentile of the population.
   (c) An 80% confidence interval will in general be narrower than a 60% confidence interval.

Properties of the Bootstrap

In the bootstrap, we have a sample $\{X_1, \ldots, X_n\}$ from which we sample with replacement n times to obtain $\left\{\widetilde{X}_1, \ldots, \widetilde{X}_n\right\}$. Most likely, some of the values $\{X_1, \ldots, X_n\}$ will show up more than once.

10. For a really big sample, how likely are we to observe a data point X_1 in a particular bootstrap sample? Write down a guess.

11. Let's see how we would answer this question analytically. First, pick a fixed sample size n. What is the probability that X_1 appears on the second draw of a bootstrap sample?

12. What is the probability that X_1 appears in a particular bootstrap sample?

13. What is the limit of this probability as n approaches ∞?

    Hint: Define $y = \mathbb{P}\left(X_1 \text{ doesn't appear in the bootstrap sample}\right)$. Then take the natural log of both sides.

14. Approximately what is the limit above equal to numerically?

15. How many times does a data point X_1 show up on average in the bootstrap sample?