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:

   Solution: A set of assumptions about the process that generated data. Some examples of assumptions include

   - The data are drawn independently.
   - The data are uniformly distributed
   - There are two populations present in the data

   (b) Null Hypothesis:

   Solution: The null hypothesis is a statement about the model that corresponds to the idea that any observed difference is due to sampling or experimental error. We are often trying to disprove this.

   (c) Test Statistic:

   Solution: A statistic (a function of the data) that can be used to help reject or fail to reject the null hypothesis.

   (d) Sampling distribution

   Solution: The distribution of all the possible values of a statistic with a fixed sample size. The assumptions of the null model should specify a sampling distribution.

   (e) p-value:

   Solution: The chance, under the null hypothesis, of getting a test statistic equal to or more extreme than the observed test statistic.

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.

   > **Solution:** True, the p-value is in the context of the null model (ie. it is the probability of extreme data under the null model)

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

   > **Solution:** False, the p-value is the probability of extreme data given the null hypothesis.

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

   > **Solution:** False. If we get a small p-value, we think that the evidence is strong enough to reject the null hypothesis, i.e. we no longer think random chance in a null model is an adequate explanation for the variability.

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

   > **Solution:** False, the p-value is the probability of observing the data under all the assumptions of the null hypothesis.

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

   > **Solution:** True; p-values are often misused/misrepresented through p-hacking and multiple testing without appropriate correction. Therefore, more information about the hypothesis testing procedure is need to understand the evidence regarding a model.

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:

1	2	2	3	3	3	4	4	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.

   Solution: $\mathbb{E}$ Single draw $=$ Average of the Box $= \dfrac{\text{Sum of the tickets}}{\text{Number of Tickets}} = \dfrac{27}{9} = 3$

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

   Solution: $\mathbb{E}\ \underbrace{\text{Average of nine draws}}_{\text{Estimator}} = \underbrace{\text{Average of the Box}}_{\text{Bootstrap population Parameter}} = 3$

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

   Solution: $\frac{93}{9} - 3^2 = \frac{4}{3}$

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

   Solution: $\frac{4}{3}$

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

   Solution: $\dfrac{\frac{4}{3}}{9} = \dfrac{4}{27}$

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

> **Solution:** We cannot calculate this from the sample. It can be estimated by the number in part e.

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.

> **Solution:**
>
> 1. Draw a bootstrap sample of size 9 from the box (bootstrap population)
> 2. Calculate the mean of the bootstrap sample.
> 3. Steps 1 and 2 constitute a single bootstrap replicate. Repeat them a large number of times(10000 is usually suggested).
> 4. Calculate the variance of the means from step 2. This is the bootstrap estimate of the population mean.

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

> **Solution:**
>
> - Estimation error - from estimating using a sample rather than direct calculation using the population distribution function (we don't have this!)
> - Simulation error - from simulating the bootstrap sampling distribution. We can reduce this by increasing the number of bootstrap replications we do or by enumerating all possible bootstrap resamples.

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

   Solution: No, 6 is not part of the original sample

   (b) 1, 2, 2, 2, 3, 3, 3, 4, 4, 5

   Solution: No, this resample size (10) is too big

   (c) 1, 1, 1, 1, 1, 1, 1, 1, 1

   Solution: Yes, this could be a resample

   (d) 2, 2, 3, 3, 3, 4, 4, 4

   Solution: No, this resample size (8) is too small

   (e) 1, 2, 3, 3, 3, 4, 4, 5, 5

   Solution: Yes, this could be a resample

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

Solution:

- The sample is representative of the population (drawn from the same distribution and is "big enough")
- The sample was drawn i.i.d.

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.

Solution: The medians for the bootstrap resamples are:

$$2, 3, 3, 1, 3, 3, 3, 3, 4, 2$$

Sorting these values:

$$1, 2, 2, 2, 3, 3, 3, 3, 3, 4$$

We take the 20^{th} and 80^{th} percentiles to be the endpoints of our confidence interval, giving us [2, 3].

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.

   Solution: No, the confidence interval either covers the true population parameter or it doesn't. The 60% refers to the coverage of the many confidence intervals constructed from hypothetical new samples.

   (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.

   Solution: This statement is fine. 40% intervals are contained in 60% intervals.

   (c) An 80% confidence interval will in general be narrower than a 60% confidence interval.

   Solution: No, they are wider.

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?

    Solution: $\frac{1}{n}$

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

    Solution: The probability that we don't pick X_1 in the ith draw is $1 - \frac{1}{n}$. The probability that we don't pick X_1 at all is the probability that we don't pick it for the 1st draw, for the second draw, and so on up to the nth draw. Since we are sampling them independently, this probability is the product

    $$\mathbb{P}\left(X_1 \text{ doesn't appear in the bootstrap sample}\right) = \underbrace{\left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{1}{n}\right)}_{n \text{ times}} = \left(1 - \frac{1}{n}\right)^n$$

    This is the probability that we don't observe X_1 at all. We are interested in the probability of the complement event: observe X_1 at least once. This is given by one minus the above probability:

    $$\mathbb{P}\left(X_1 \text{ appears at least once in the bootstrap sample}\right) = 1 - \left(1 - \frac{1}{n}\right)^n.$$

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.

Solution: Define $y = \left(1 - \frac{1}{n}\right)^n$ and take the log of both sides to get

$$\ln y = n \ln\left(1 - \frac{1}{n}\right) = \frac{\ln\left(1 - \frac{1}{n}\right)}{\frac{1}{n}}$$

Take the limit of both sides and apply L'Hôpital's Rule.

$$\lim_{n\to\infty} \ln y = \lim_{n\to\infty} \frac{\ln\left(1 - \frac{1}{n}\right)}{\frac{1}{n}} = \lim_{n\to\infty} \frac{-\left(\cancel{-\frac{1}{n^2}}\right)\frac{1}{1-\frac{1}{n}}}{\cancel{-\frac{1}{n^2}}} = \lim_{n\to\infty} -\frac{1}{1-\frac{1}{n}} = -1$$

Exponentiating both sides,

$$e^{\lim_{n\to\infty} \ln y} = \lim_{n\to\infty} e^{\ln y} = \lim_{n\to\infty} y = e^{-1}$$

Putting things together:

$$\lim_{n\to\infty} \mathbb{P}\left(X_1 \text{ appears at least once in the bootstrap sample}\right) = 1 - e^{-1}$$

Technical points: There is a minor abuse of notation when we take the derivatives since $n \in \mathbb{N}$. Here you should understand the numerator and denominator as continuous functions of real numbers. For those of you who are worrying about the existence of this limit, you can be more careful using a squeeze theorem argument. See `http://www.maths.manchester.ac.uk/~mprest/elimit.pdf` for a closely related proof.

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

Solution:

$$\lim_{n\to\infty} \mathbb{P}\left(X_1 \text{ appears at least once in the bootstrap sample}\right) = 1 - e^{-1} \approx 0.632$$

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

Solution: First note that the number of times T that X_1 is picked for the bootstrap sample can be written as

$$T = \sum_{j=1}^{n} I[\widetilde{X}_j = X_1],$$

where $I[\widetilde{X}_j = X_1]$ is 1 if $\widetilde{X}_j = X_1$ and 0 if $\widetilde{X}_j \neq X_1$. To get the expected number of times, use linearity of expectation!

$$\mathbb{E}[T] = \sum_{j=1}^{n} \mathbb{E}I[\widetilde{X}_j = X_1] = \sum_{j=1}^{n} \mathbb{P}\left(\widetilde{X}_j = X_1\right) = \sum_{j=1}^{n} \frac{1}{n} = 1.$$

So the expected number of times X_1 shows up is 1, and the same is true for every other observation $X_2, \ldots, X_n$. This makes sense, since there are n candidate positions and n observations to choose from, with no one privileged over the other.