IB120/201 - Lab 9

Bootstrapping

Due Date: March 27, 2020

University of California, Berkeley *GSI: Naveed Ziari*

In this lab, we will learn how to extrapolate important parameters and numerical values for statistical inference from a dataset. In addition, we will learn about and implement the bootstrap. These are fundamental tools required in the analysis of large datasets in computational biology.

Background

Sampling Distribution

Statistical analysis involves making inferences about a population based on a certain sample of that population. For instance, assume that we want to find out a certain attribute of a large population (e.g. average IQ of the entire state of California). It is impossible to obtain observations from every resident of California, so we rely on samples of that population to make inferences. If we take repeated samples out of the population, and calculate the mean of each, we get a *sampling distribution of the sample mean*. This means that making an inference about the mean of the population by sampling resembles a probability distribution. The central limit theorem states that the sampling distribution of a population resembles a normal distribution even when the population itself is not normally distributed.

Bootstrap

Bootstrapping is a general class of methods that involve sampling *with replacement*. Going back to our example above, suppose we were able to collect data from N people. This sample of size N has only one mean (and median, variance, etc.), but we are interested in a distribution to obtain confidence intervals and to see how well the sample generalizes across the entire population. Bootstrapping entails *resampling* from the sample of size N *with replacement* N times to create a new "bootstrap sample" of size N. The bootstrap samples therefore have the same size as the original sample but may have repeated values and hence missing some values in the original dataset. This procedure is repeated k number of times (k is typically a large value). From these bootstrap samples, we now have a distribution of bootstrap sample means. Bootstrapping is a powerful technique because it allows us to calculate the uncertainty of estimators (which in this case is just the mean).

Questions

1. Load the `iris` dataset, and find the mean and SEM of the `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`.
2. Calculate the variance of `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`.
3. Using the values calculated above, obtain the 95% confidence interval of the mean for `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`. What determines the range of the confidence interval?
4. Obtain 500 bootstrap samples out of `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`:
   (a) Create a histogram of the bootstrap sample means
   (b) Find the 95% bootstrap confidence intervals

(c) Write a function for your own estimator (something other than the mean - could be for instance the median or skew) and compute the 95% confidence interval for it.

5. Perform a t-test between all combinations of the four datasets (that would be a total of 4 choose 2 = 6 combinations). See if there is any statistically significant relationship between any of the four parameters defining the `iris`.