**CSCE 623 Spring 2020: Machine Learning In Class Work, Day 9**

From Chapter 5: Resampling / Bootstrapping

Bootstrapping code practice:

Goal: to show how bootstrapping can be used to generate estimates of the mean and variance of a population with a small subset sample of data. Our target distribution is discrete distribution of a set of integer sums of 3 fair 6-sided dice rolls (3d6). The range of resulting integers is 3-18, and will approach a discretized version of a bell curve (if you ignore the missing infinite tails)

Overview:

Task 1: Determine the true population discrete distribution, mean, and variance of an infinite number of 3d6 dice rolls

Task 2: Compute estimates on the population using a small sample from the population and bootsrapping.

Task 3: Compute estimates on the population using a large pseudorandom sample from the population directly.

Numpy.random functions such as choice, randint, shuffle, or permutation might be useful.

**Task 1: determine the true population mean and true population variance from an infinite number of 3d6 rolls:**

A. Generate the set of all possible sums from dice rolls of 3d6. One way to do this is with an array (or list) of size 6 x 6 x 6 = 216 values. Each member should be an integer on the interval [3, 18]. A histogram with integer-width bins should show something that looks close to a discrete bell curve. This represents the true population of possible dice roll outcomes.

B. Compute the mean of the dice rolls from step A. This is the mean of the true population

C. Compute the variance of the 216 dice rolls:  This is the variance of the true population. Compute the standard deviation of this population from the variance.

D. Using python, generate a (bar) histogram of the values from A to report the counts of dicerolls in each bin from 3 through 18. Add a black vertical line at the mean of the distribution (computed in step B). Add black dashed lines at +/- 1, 2, and 3 standard deviations from the mean (computed in step C). Confirm that your histogram approximates a discretized version of this distribution:

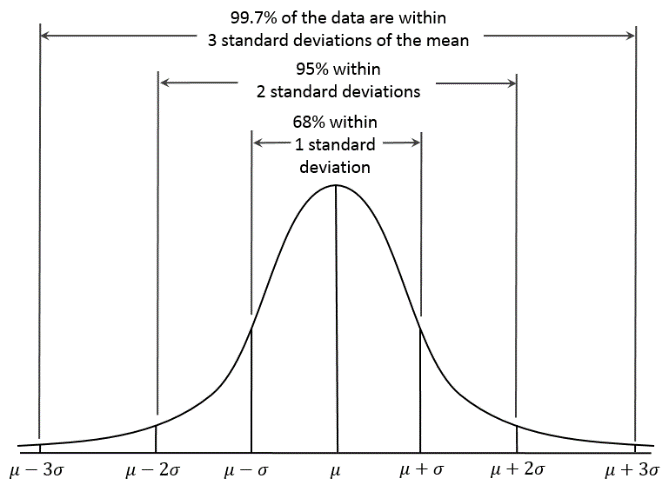


Image By Dan Kernler - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36506025

When you are convinced this is working, proceed to the next page

**Task 2: Execute the following procedure for using bootstrapping to estimating the mean and variance on the distribution of sum of rolls of a set of 3 fair 6-sided dice:**

For this part if you don’t have a six-sided die, then generate pseudorandom die rolls with an online tool or app for dice rolling or generate a random uniform sample integers on [1,6] using python code.

1. Generate the sample data using a true random source & bootstrapping:

1a. Roll 3 x 6-sided dice and add their values together (3d6). Record the resulting integer. Repeat this 3d6 rolling & summation process 7 times to obtain a total of 7 integer sums. Each of your 7 sums should be in the closed interval [3,18]. Note that you will have fewer than all the possible values in this interval, and you may have repetitions of some values.

1b. Store these 7 sums in a 1x7 (column) vector in python. This vector is one observation.

1c. Make a histogram of these 7 values. What can you say about the shape of the histogram? Is it what you expected? At least 16-7=9 possible values will be missing from this set of dicerolls.

1d. A bootstrap is an instance of **resampling with replacement**. The size of a bootstrap sample is the same size as the original data – in this case, 7. Using the 7 integer-vector from step 1b as your fixed starting point, create 1000 bootstrap resamples of those 7 integers. Store your result will be a 1000 row x 7 column matrix where each row is a bootstrapped sample and each cell of that row is an integer (possible result of a 3d6). Note that if you never rolled a value (e.g. 18) in the original 7 rolls, it will never appear in this bootstrap sample either.

2. Compute the mean and variance of each bootstrapped sample (row). You will have 1000 estimates for mean and 1000 estimates for variance. Try to do this without using a for loop via matrix multiplies and/or sums.

3. Using the formula on page 188 where the αr are the means from step 2, Compute the estimate of the mean of the population from the bootstrapped population

4. Similarly, using the second equation on page 188 and the result from step 3 and the αr means to compute the estimate of the overall standard deviation from the bootstrapped population, and variance

5. Make a histogram of the 1000 *means* from step 2 and add a red vertical line at the estimated population mean from step 3 and the actual population mean (from pre-task step B) with a blue vertical line. How close were you?

6. Make a histogram of the 1000 *variances* from step 4. Add a red line at the estimated population variance from step 4 and mark the actual population variance (from pre-task step C) with a blue vertical line. How close were you?

**Task 3: Obtaining statistics from actual (simulated) draws of the distribution:**

1. Generate the sample data using a pseudorandom source and simulation: Simulate and store the result of 1000 repetitions of generating 10 instances of the integer sum of 3 rolls of a fair 6-sided die   
(each of the 1000 rows will have 10 columns of 3d6). The result of this will be a matrix (size 1000 x 10) of integers. Each row is a *sample* of 10 integers from the population of possible 3d6 dice rolls. Ensure all values of elements in this matrix are in the closed interval [3,18]

2. Compute the sample mean and variance of each sample of 5 integers. (you will have 1000 estimates of mean & 1000 estimates of variance). Try to do this without using a for loop.

3. Using the formula on page 188, Compute the overall mean of these 1000 means.

4. Using the formula on page 188, Compute the overall standard deviation of these 1000 means, and compute the variance of the 1000 means from the standard deviation.

5. Make a histogram of the *means* of these 1000 statistical values. Mark the sample estimate for population mean with a red vertical line, and mark the actual population mean (from pre-task step B) with a blue vertical line. How close were you?

6. Make a histogram of the *variances* of these 1000 statistical values. Mark the sample estimate for population variance with a red vertical line, and the actual population variance (from pre-task step C) with a blue vertical line. How close were you?

**Now compare the results of the bootstrapped (task 2) vs. repeated sampling from the true population (task 3)**