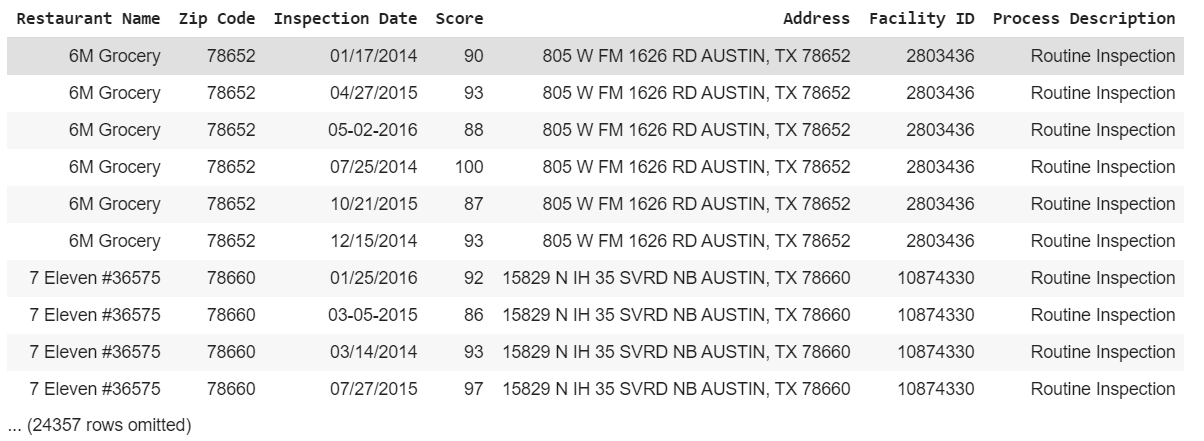
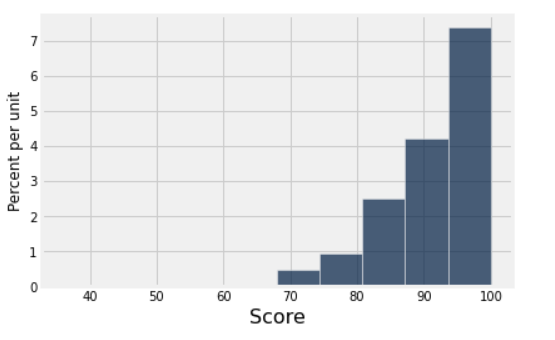
Aditya Pitchuka  
INFO 5502.002 (14676)  
Spring 2020

**Assignment – 6**

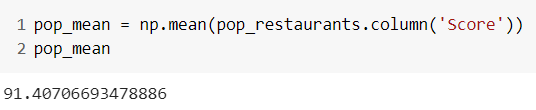
Reading the given dataset.

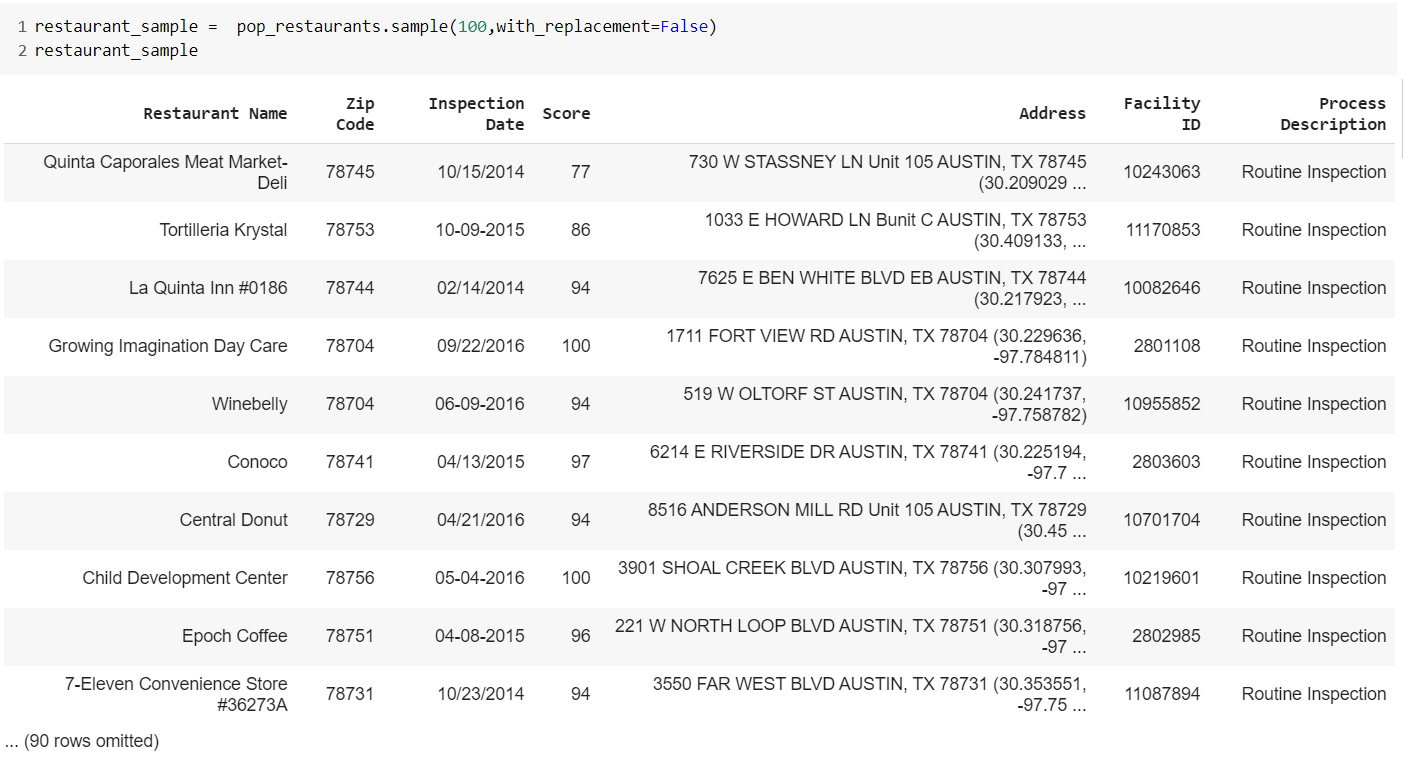


Plotting the histogram for Restaurant’s score.

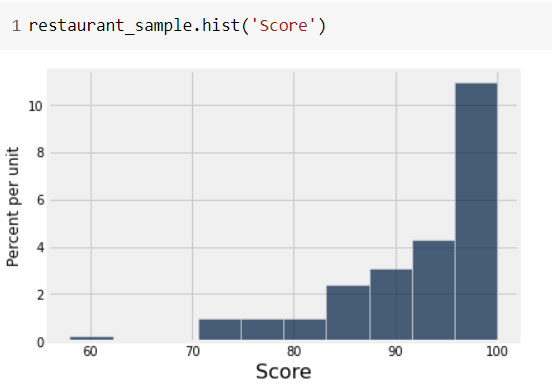


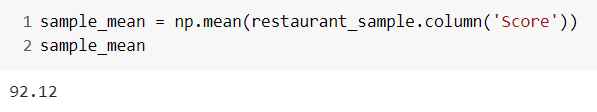
Finding the mean of scores of the given population.



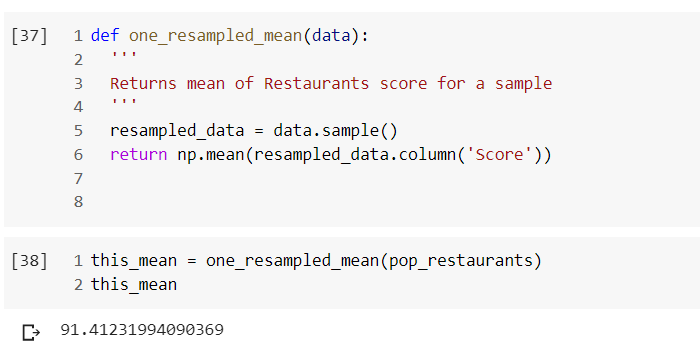
Making a sample, without replacement, of size 100. 

Now plotting the histogram for score for the sample obtained and finding its mean.

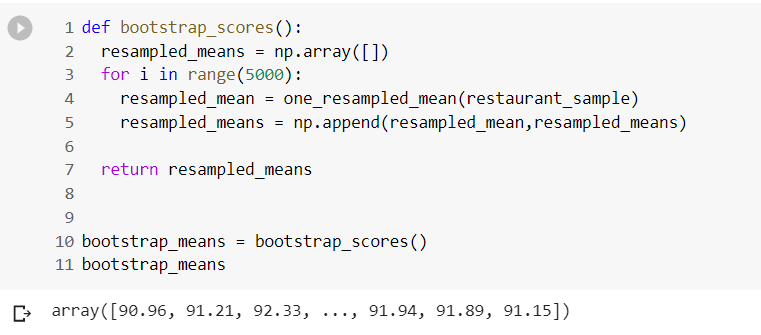




**Q 1.1** Complete the function one resampled mean below. It should take in an original table data, with a column Score, and return the mean score of one resampling from data.

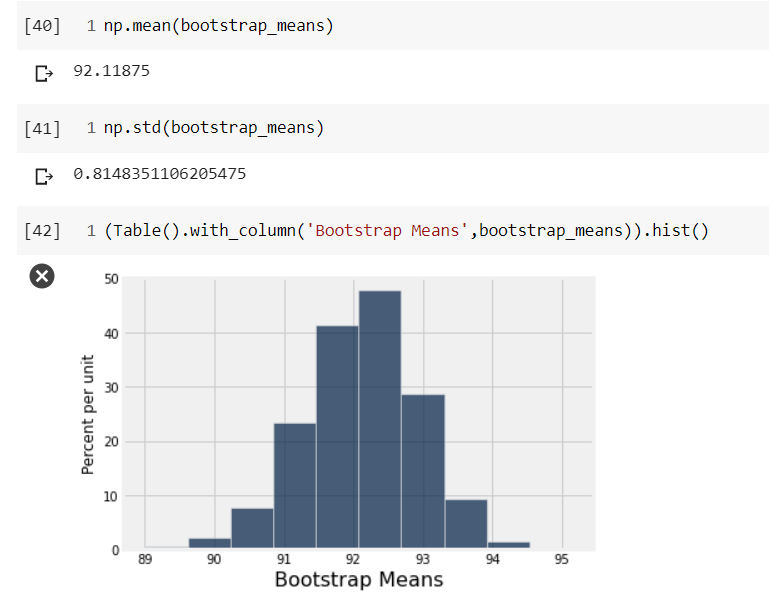


**Q 1.2** Complete the function bootstrap\_scores below. It should take no arguments. It should simulate drawing 5000 resamples from restaurant sample and compute the mean restaurant score in each resample. It should return an array of those 5000 resample means.



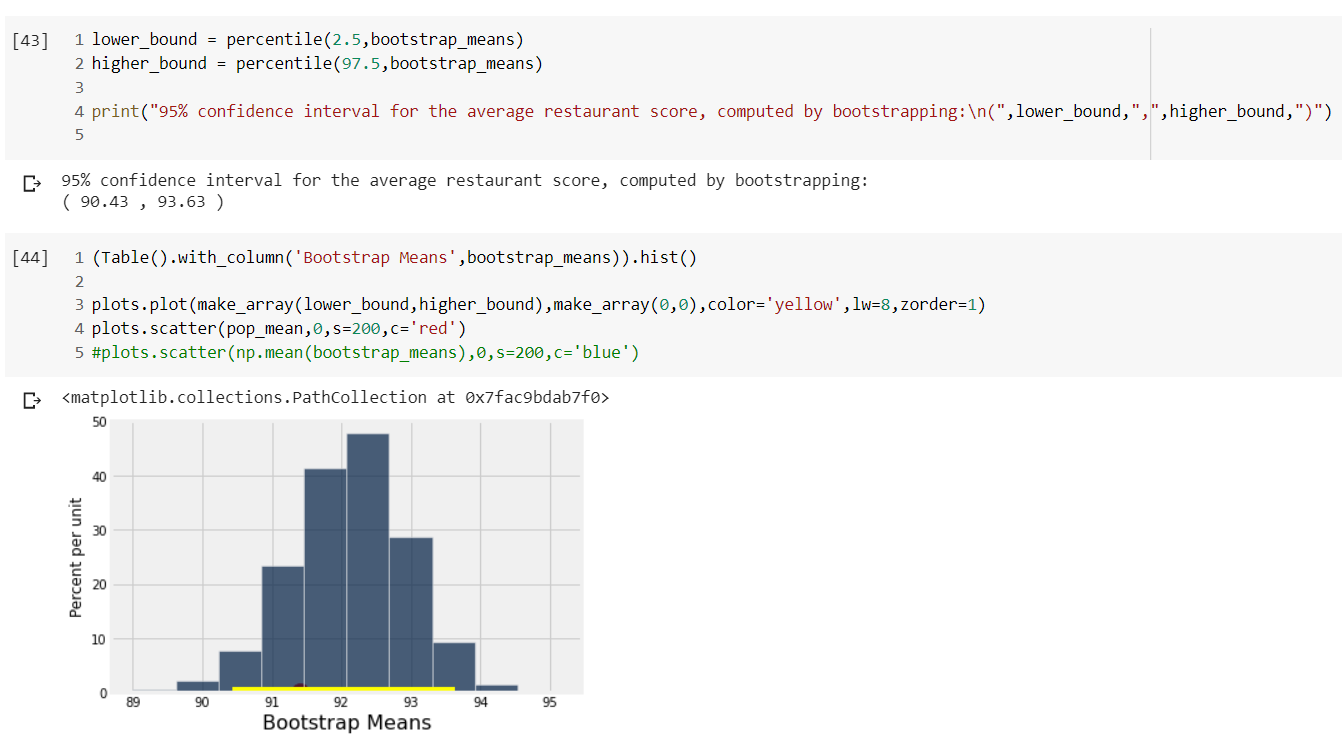
Finding the mean and standard deviation of the sampled means of the scores and plotting a histogram for the obtained array of samples of means.

We observe that the distribution looks like a normal distribution even though the original sample and the original population is far from normal distribution.



**Q 1.3** Compute a 95 percent confidence interval for the average restaurant score using the array resampled means.

Also plotting a histogram of the distribution with the confidence interval, in yellow line, and the population mean pointed in red dot.



**Q 1.4** What distribution is the histogram between question 2 and 3 displaying (that is, what data are plotted), and why does it have that shape?

In question 1.2, all the sampled means are calculated and plotted in a histogram to give us an idea about the variablilty of sample means of the different samples drawn from a population, but in question 1.3 we are doing it using only a single sample. We are doing resampling 5000 times as the law of large numbers state that when an experiment is repeated large number of times the estimate will be close to the real value. And when the sample is large enough the real value will be close to population parameter.

**Q 1.5** Does the distribution of the sampled scores look normally distributed? State \yes" or \no" and describe in one sentence why you should expect this result.

Yes, the distribution of the means looks much closer to a normal distribution. The distribution seems more and more into an empirical distribution the more we simulate many many times.

**Q 1.6** Without referencing the array resampled means or performing any new simulations, calculate an interval around the sample mean that covers approximately 95% of the numbers in the resampled means array. This confidence interval should look very similar to the one you computed in Question 3.



We used the central limit theorem. Which states –

“*The sampling distribution of the sample mean is approximately normal for large sample sizes, regardless of the distribution of the population.”*

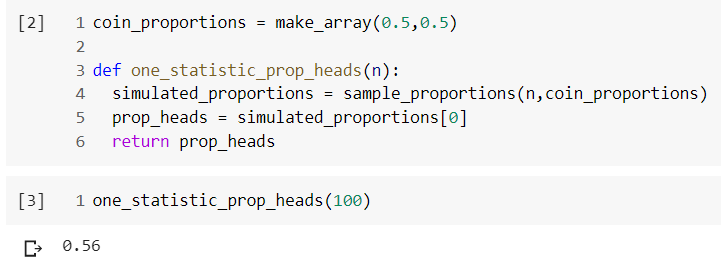
For the random samples we take from the population, we can compute the mean of the sample means:



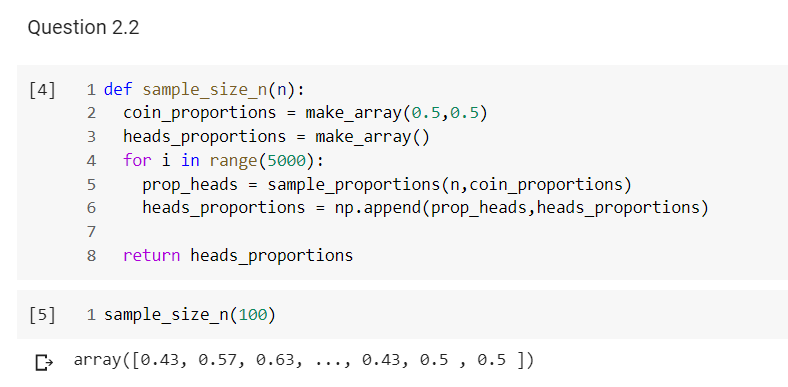
and the standard deviation of the sample means:



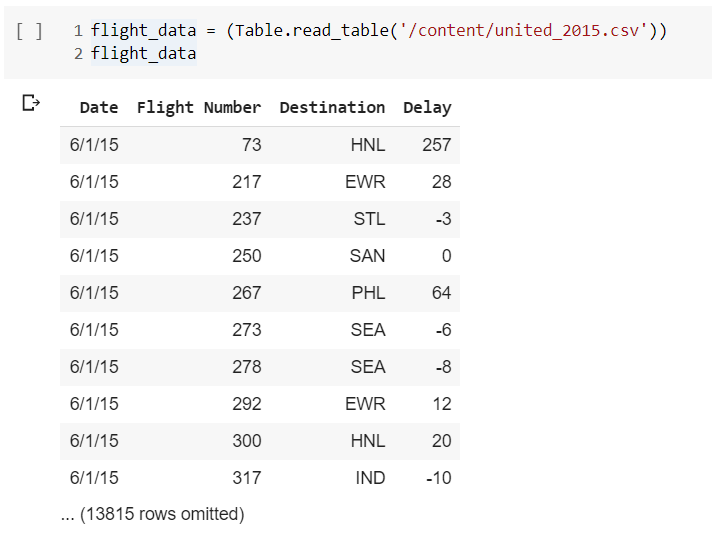
**Q 2.1** Define the function one\_statistic\_prop\_heads which should return exactly one simulated statistic of the proportion of heads from n coin flips.



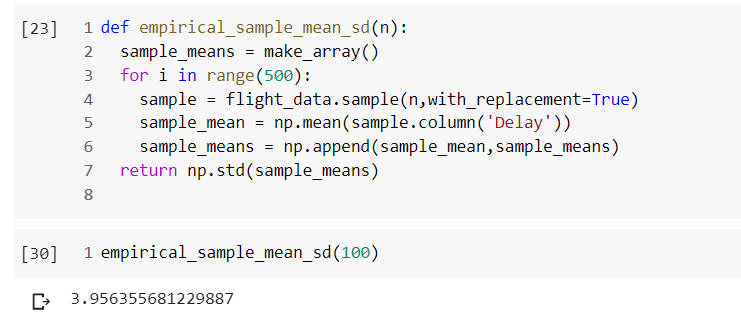
**Q 2.2** Write a function called sample\_size\_n that takes in a sample size n. It should return an array that contains 5000 sample proportions of heads, each from n coin flips.



Using the flight database to test the central limit theorem. Reading the dataset.



**Q 2.3** Write a function called empirical\_sample\_mean\_sd that takes a sample size n as its argument. The function should simulate 500 samples with replacement of size n from the flight delays dataset, and it should return the standard deviation of the means of those 500 samples.



**Q 2.4** Now, write a function called predict\_sample\_mean\_sd to find the predicted value of the standard deviation of means according to the relationship between the standard deviation of the sample mean and sample size that is discussed in the textbook. It takes a sample size n (a number) as its argument. It returns the predicted value of the standard deviation of the mean delay time for samples of size n from the flight delays (represented in the table united).

