Fundamentals of Data Science

Coding assignment - spring 2023

We are given the dataset of weights of newly born babies. There are 10 data files. According to the data allotment scheme, data6 assigned us to do this coding assignment. There are 400 data records and just a single data column named weight.

The histogram of given data comes up with a well-defined bell curve shape. Our data is not perfectly normally distributed but approximately well normally distributed. The mean value of data is 3.40 with a standard deviation of 0.38. The maximum value in the data is 4.47314, and the lowest value is 2.35406.

How do you calculate the mean weight? What value do you get?

To calculate mean weight we utilized the mean function of numpy(numpy.mean(data\_values)

How do you calculate the required value X? What value do you get?

Basically, we tried two methods to calculate X.

* Z Score Method
* Percentile Method

Z Score Method:

After plotting a histogram of the given data, we found that the data under analysis is well normally distributed which means we can apply the Z Score method to find the value of X. As we directed to calculate the value of X such that 10% of newborns from the distribution are born with a weight above X which corresponds to Z=1.28.

X=mean + standard\_deviation \* Z\_Score

Here X represents the value that is less than 10% of the data.

Percentile Methods:

In percentile, we divide data into 100 equal parts. Our target is to filter out values greater than the 90th percentile, which means only 10% of values greater than 90%, which is our target value.

X = (n/N) × 100

n is the nth percentile

N is the total number of values

X is the value at nth percentile