Homework 10: Statistics and scientific reproducibility

December 7, 2021

1

1.a

I would first calculate the mean value from the data, and then derive the standard deviation from the mean. The standard deviation can be considered to be a measure of how dispersed the data is from the mean, and thus a low standard deviation would mean that values are clustered close to the mean and high standard deviation would indicate data that is more spread out. Thus, the mean together with the standard deviation would be a valid presentation of the result and the accuracy of the result.

1.b

The first data set has a range of approximately $13.3 - 5.3 = 8$, while the second data set have a range of approximately $14.9 - 3.9 = 11$. This indicates that the second data set has a larger spread of values, and would probably have a larger standard deviation compared to the first data set. This is not a significantly large difference, but the mean value in the first data set would probably be more accurate than in the second.

1.c

The mean derived from this data set is 12.47. However, the median is as low as 7.58. This can be explained by inspecting Figure 1. As we can see, the majority of values are between 0 and 25, but there's a minority of values between 25-200 that increases the mean. Considering this, the median value would probably represent this data set more accurately.

When it comes to calculating the confidence interval for the median value, the upper bound is defined as $n \cdot q + \sqrt{n \cdot q \cdot (1-q)}$ and lower bound as $n \cdot q - \sqrt{n \cdot q \cdot (1-q)}$, where n is the sample size, q is the quantile of interest, and z is the z-score. For median q is usually chosen as $q = 0.5$, and z-score is dependent on the chosen confidence interval. For a confidence level of 0.95 for example z should be $z = 1.96$.

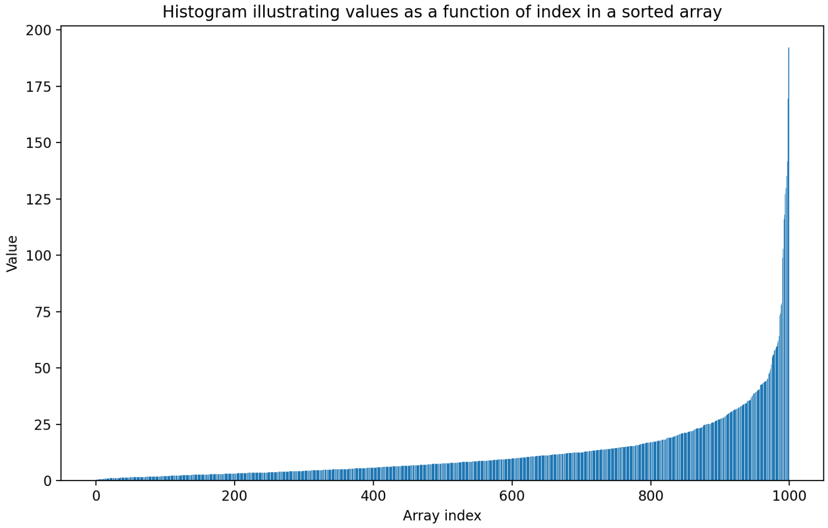

Figure 1: The histogram illustrates the values of the data set in a sorted array, as a function of the array index.

2 Reproducability of BERT paper

I have chosen to take a closer look at the reproducability of the BERT paper, written by Devlin et al. at Google. The paper focuses mainly on two phases, pre-training and fine-tuning. I would argue that the information included in the paper isn't sufficient to reproduce the pre-training phase. Moreover, the computational resources required to pre-train BERT are vast, which could be considered a drawback, since fewer institutions will have the means to reproduce the results. However, Devlin et al. does reference an already pre-trained model that can be used when conducting similar experiments. The fine-tuning phase however is described in sufficiently detailed manner to reproduce the experiments. Furthermore, the fine-tuning phase is less computationally demanding, which lowers the barriers for reproduction. Finally, all of the code for BERT is available on github and referenced in the paper, which additionally lowers the barriers of reproduction. [1]