Computational Methods for Statistics

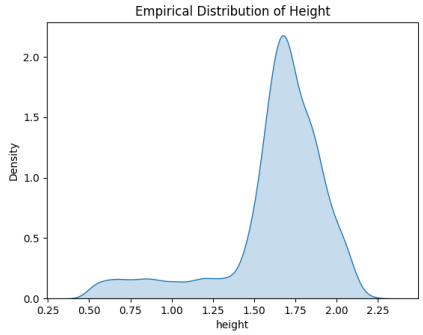
Saimir Guda
Matriculation Number: 11933108

January 2023

1 Bootstrap and confidence interval

1. A social science student collected data on the height of Austrian people in a file height.csv and asks you to do an analysis of this dataset in order to publish the results in a well known science magazine. The data set contains the height of male and female persons i.e. a column which indicates the respective height and gender of each person. Note that also a certain amount of kids is included into the data.

   (a) Visualize the empirical distribution of all the heights recorded in the data.



   (b) Now get a more detailed insight into the data by computing some test statistics which measure various properties of the dataset. The test statistics of choice are the median, IQR, 90th percentile, the absolute difference of mean and median, the skew and the kurtosis. Explain to the social scientist (i.e. include in your report) what, in a practical sense, is measured by the various quantities. For instance you would mention that the mean height is a measure for the average size of a person. But you would also point out that only a few outliers attaining extreme values can shift the mean drastically, which is something to look out for. In that particular setting, how could the affect of children be interpreted on the various test statistics?

      i. MEDIAN: 1.68703
      The median is the value exactly in the middle of the sorted value of arrays. this means that more than half of the data values are lower than 1.68703 m.

      ii. IRQ: 0.25385
      The interquartile range is a measure of where the “middle fifty” is in a data set. Where a range is a measure of where the beginning

and end are in a set, an interquartile range is a measure of where the bulk of the values lie. The higher the IQR, the more dispersed the are the measurements in this (middle) range. An IQR of 0.2538 indicates a not so a not so enormous scattering of the body sizes in the middle 50

iii. 90th PERCENTILE: 1.9385
This value indicates the threshhold at which lie 90% of the samples. Every value from 1.9385 and higher would belong to the other 10%.

iv. ABS diference: 0.0615
It indicates outliers in the data set such as child sizes.

v. SKEW: -1.58349
Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. This basicly means that there are less concentrated measurementsin the plot on the left side(so child messurements), and that the mass of measurements is on the right(all other adults).

vi. KURTOSIS: 2.458
The kurtosis indicates how far the margins of a distribution deviate from the normal distribution, or how steep the curve is. It is positive so it is a steep curve.