(**Bootstrap confidence interval: 10+20 points**) 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) First of all, get an impression of the data. To that end, 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?

(c) Now compute confidence intervals for each test-statistic you evaluated. This will improve the quality of your work and inform the readership about the range of statistical error. Since the distribution of some of these test-statistics can be very complicated, we use the bootstrap method to come up with confidence intervals. Draw $b = 1000$ bootstrap samples of the height observations and based on this, compute the 90% pivotal confidence intervals for each respective measure. Visualize the respective bootstrap distributions, each one in a separate plot of a 2×3 tile of plots. In each plot draw a vertical line which indicate the realization of the test-statistic on the data and mark the confidence intervals. What changes if instead of the pivotal method, you use the percentile method?

(d) One interesting question might be, whether the height of male and female are different, based on that dataset. Compute the Pearson correlation coefficient between the height of male and female. In order to again provide confidence intervals, use $b = 1000$ bootstrap samples. Make sure you preserve the pairings of gender and height by only drawing the row-indices. Again use the pivotal method to compute a 90% bootstrap confidence interval and visualize the findings as in task (c). Also show the simulated variability of the test-statistic around the observed value by drawing the empirical distribution of the bootstrap error $\delta_n^* = T_n^* - t_n$.

(e) Repeat tasks (c) and (d) with only a tenth of the bootstrap samples and once with ten times as many bootstrap samples. Which differences do you observe? Would the assumption that men are in general larger than women be supported by your analysis?