Computational Methods for Statistics

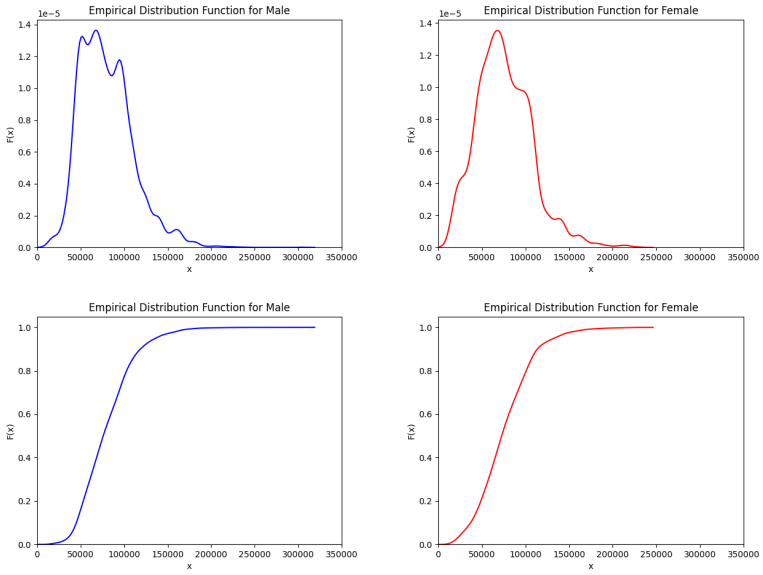
Saimir Guda
Matriculation Number: 11933108

January 2023

1 Permutation Test:

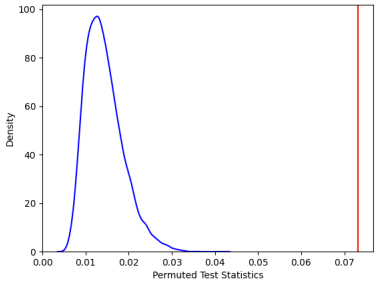
The same social scientist as in the previous task asks you to investigate a dataset on wages, since you've done such a good job on the previous task. The dataset wages.csv contains the annual salaries of male and female persons and a column which indicates the respective salary and gender.

a) Plot the respective empirical distribution functions of female and male people, as well as their empirical cumulative distribution function.



b) Now you perform a permutation test which evaluates 10.000 permutations to investigate the inter- changeability of data between the two groups. For each permutation you evaluate the test-statistic. In the end you obtain a vector of 10.000 numbers. Get hold of the density of the test-statistic by plotting the empirical density. Within this plot, mark the realization of the test-statistic on the original data where you have not yet interchanged anything. Estimate the p-value according to the permutation test and decide if you reject the null-hypothesis of the KS-test at a significance level of 3.5%. Could you already tell the outcome of the test by looking

at the respective ecdf?



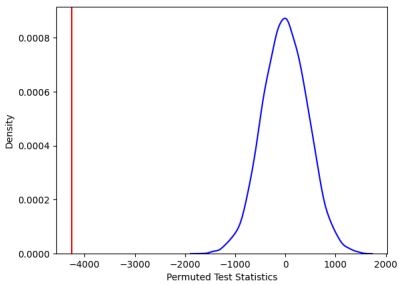
.

i. The Null hypotheses using the KS test would be that the two groups have the same distribution. The alternative hypothesis is that these two groups have different continuous distributions.

ii. The pvalue we go from the test was 0.00009999. A p-value close to 0 indicates that the observed test statistic is very unlikely to occur under the assumption that the two groups have the same distribution. This p value i also less than our significance level thus we reject the null hypothesis.

iii. From the respective ECDF graph we can indeed tell that the distribution is not the same. Parts of the graph differ quite a lot, however there is some overlapping. To have a clear answer a test must be concluded.

c) Another test-statistic which measures the difference in the wages would consider the difference in mean of the two groups. So formulate a Null-hypothesis which assumes the smaller observed mean, besides being smaller, however is not statistically significantly smaller. Using a permutation test with 10.000 permutations check, if you would reject the Null-hypothesis at a level of 1%.



.

Using the mean the pvalue of the test was 1. Since this value is above 0.01 we can accept the null hypothesis that two groups have the same distribution, however that does not mean we should. Using the mean can lead to information loss. So the information contained in the entire distribution

of the data is lost, which can lead to a loss of power in the test. Thus this method doesn't provide strong evidence toward the null hypotheses.

d) Which conclusion about fairness do you draw? Base your argument on the parts of the exercise which you solved.
The conclusion which we draw is that there is a difference between the wage distribution of male and females. This comes from a lot of different factors which are not covered in this assignment.
From looking at the ECDF and EPDF graphs from the first task we can see that the lower wages have a higher density on females and the higher wages have a higher density of men. The graph on in the middle of these 2 parts seem to balance itself and the distribution tends to get more equal, however what we can conclude from this test is that there is a pay gap between male and female people.