Machine Learning: Exercise Set IV

Georgios Manos
csd4333

9 November 2022

1 Hypothesis Testing

Starting off with the implementation details, I ran the experiments for 100 repeats and 100 permutations as I didn't have enough time for more. I'll provide all the requested plots below ($6 \cdot 3 = 18$, for each dataset and sample size), which are all labeled and describe what they represent. I also did not tamper with the histogram bins. Finally, I did not use any library functions and I implemented the CDF myself, with the formula provided by the recitation document.

1.1 Original p-value

We have the p-values table for each dataset and sample size:

	D1	D2	D3	D4	D5	D6
25	1.8110	0.5044	0.3578	0.3121	0.0003	1.8941e-05
100	0.2898	0.4138	8.0088	7.9663e-06	2.5523e-07	3.5267e-12
1000	0.0531	0.1599	0.0394	2.6065e-37	4.0164e-57	2.1529e-143

Figure 1: p-values from Chi square statistic and CDF for each dataset and sample size

If we set as acceptance threshold 0.05, we can reject null hypothesis for all cases with p-value < 0.05 and accept the others ((D3, 1000), (D4, 100), (D4, 1000), D5 and D6 columns will be rejected)

Note that the results vary a lot for different runs due to the random sampling.

1.2 Permutation Testing

The permutation testing proceedure is executed as follows:

1. for each dataset and for each sample size
    1.1. for N repeats
        1.1.1. random sample the dataset using the provided joint probability distributions and calculate chi test statistic t_o.
        1.1.2. for K permutations
            1.1.2.1. Shuffle X column, compute X^2 statistic and store the result
        1.1.3. for all permutation results, calculate the percentage of permutations where $|t_b| \geq |t_o|$, where $|t_b|$ the permutation Chi statistic test result and $|t_o|$ the original Chi statistic test
    1.2. plot p-value over N repeats

All tests were made with N = 100 and K = 100 as they were taking too long to produce. In general, we can see that for the cases we rejected the null hypothesis we usually have more $\hat{p}$-values near 0 (larger bins). The results vary a lot whenever we repeat the test, but they usually follow a similar distribution. Those distributions usually follow an increasing or decreasing pattern, except on the clearly null hypothesis rejection cases such as dataset 5 and 6. There are also some cases where they approximate a uniform distribution, which suggests that there is a high uncertainty for that case and generated dataset.

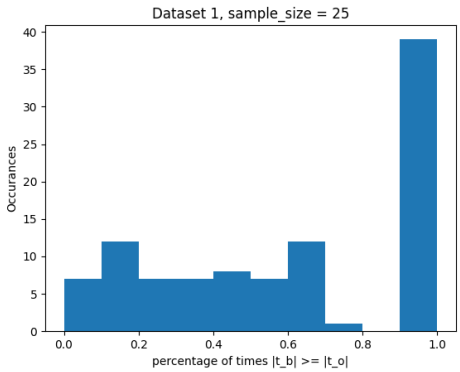

Figure 2: p-values distribution for dataset 1 and 25 samples

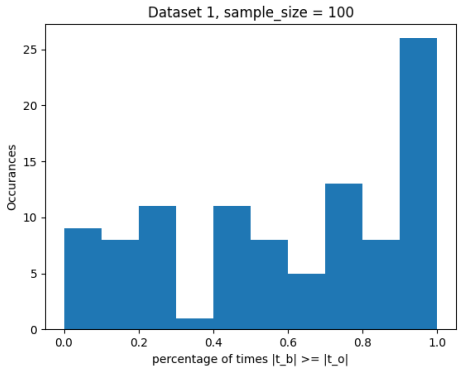

Figure 3: p-values distribution for dataset 1 and 100 samples

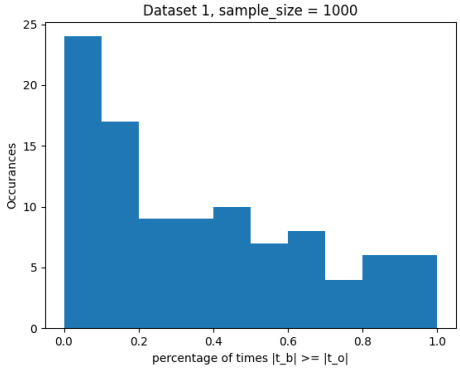

Figure 4: p-values distribution for dataset 1 and 1000 samples

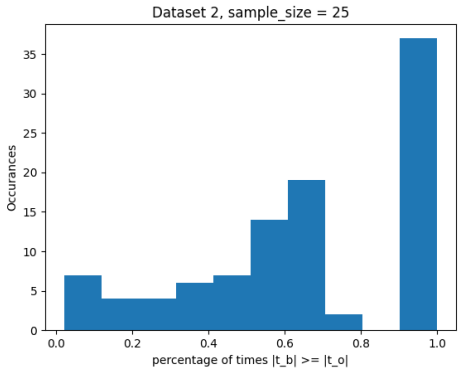

Figure 5: p-values distribution for dataset 2 and 25 samples

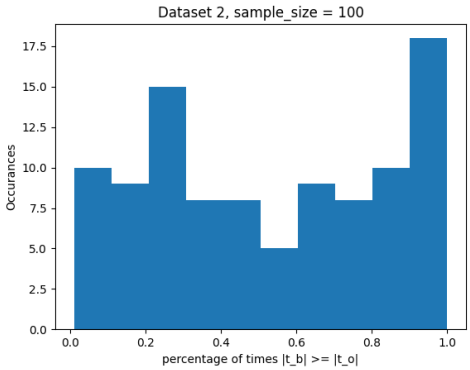

Figure 6: p-values distribution for dataset 2 and 100 samples

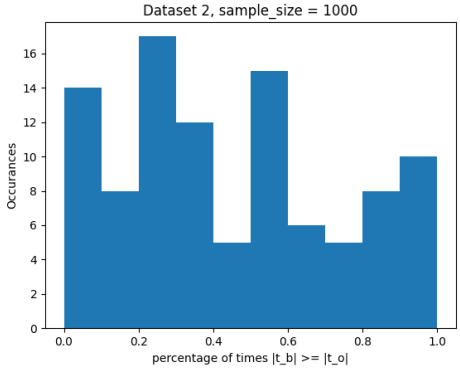

Figure 7: p-values distribution for dataset 2 and 1000 samples

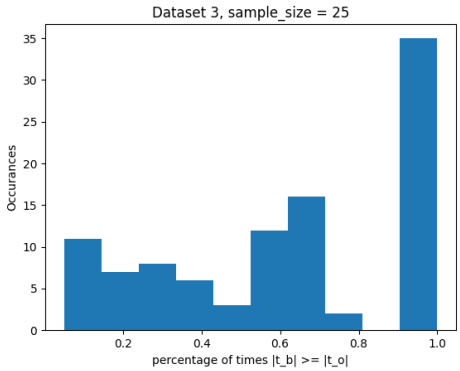

Figure 8: p-values distribution for dataset 3 and 25 samples

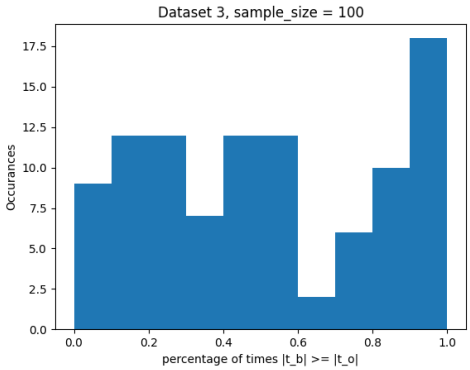

Figure 9: p-values distribution for dataset 3 and 100 samples

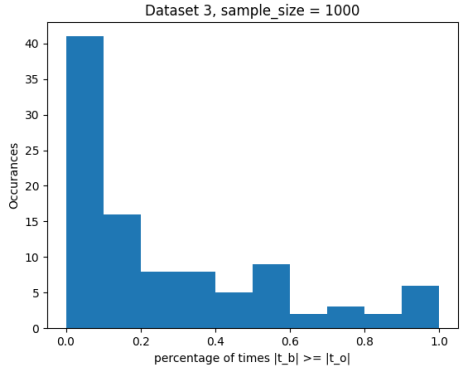

Figure 10: p-values distribution for dataset 3 and 1000 samples

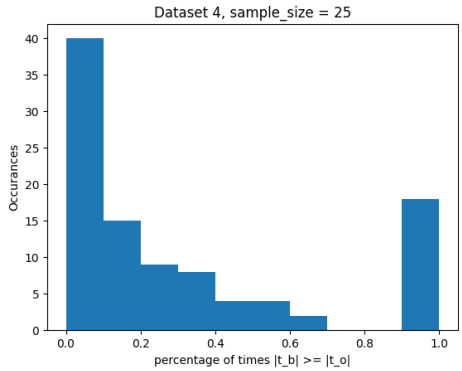

Figure 11: p-values distribution for dataset 4 and 25 samples

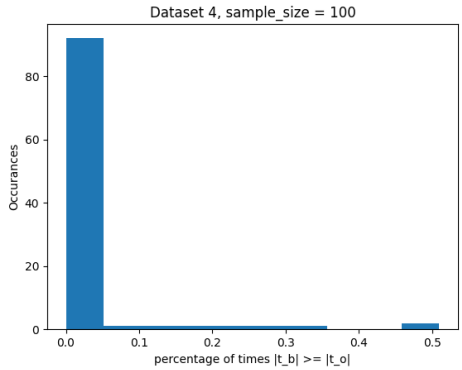

Figure 12: p-values distribution for dataset 4 and 100 samples

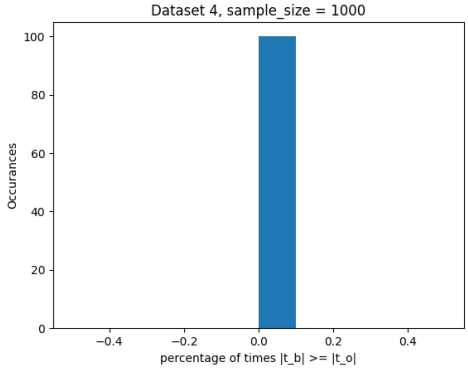

Figure 13: p-values distribution for dataset 4 and 1000 samples

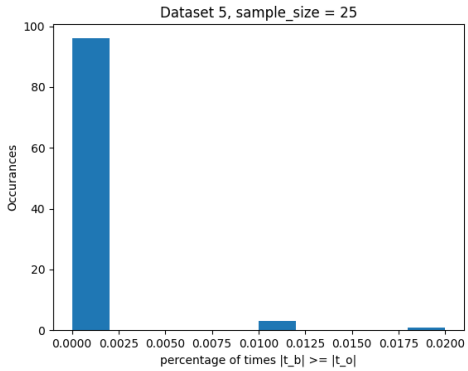

Figure 14: p-values distribution for dataset 5 and 25 samples

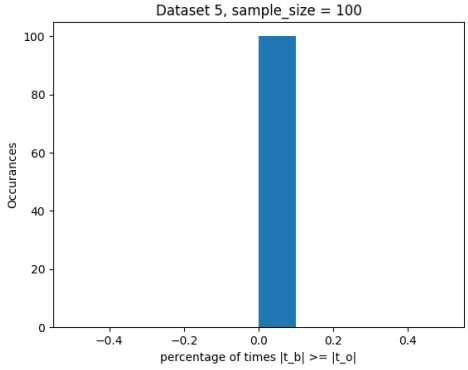

Figure 15: p-values distribution for dataset 5 and 100 samples

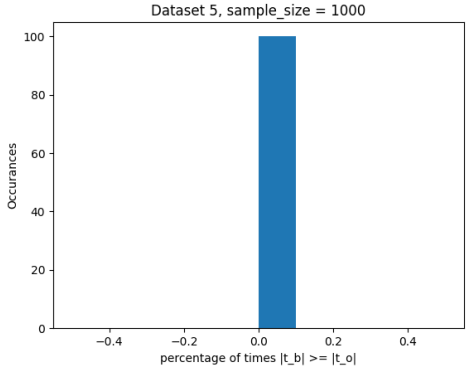

Figure 16: p-values distribution for dataset 5 and 1000 samples

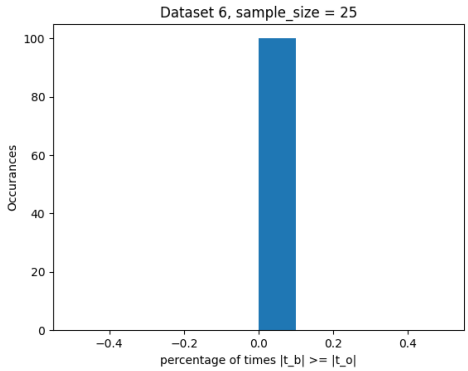

Figure 17: p-values distribution for dataset 6 and 25 samples

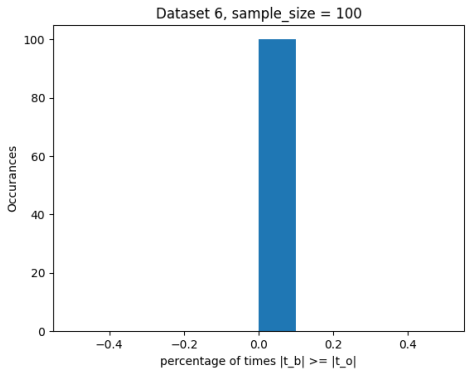

Figure 18: p-values distribution for dataset 6 and 100 samples

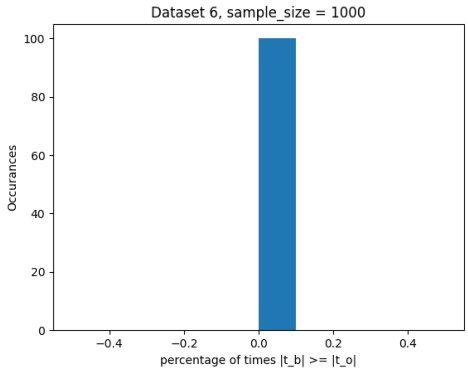

Figure 19: p-values distribution for dataset 6 and 1000 samples