Biostatistics

In-class exercise week 5: solution
Topic: multiple testing, interpretation of a p-value histogram

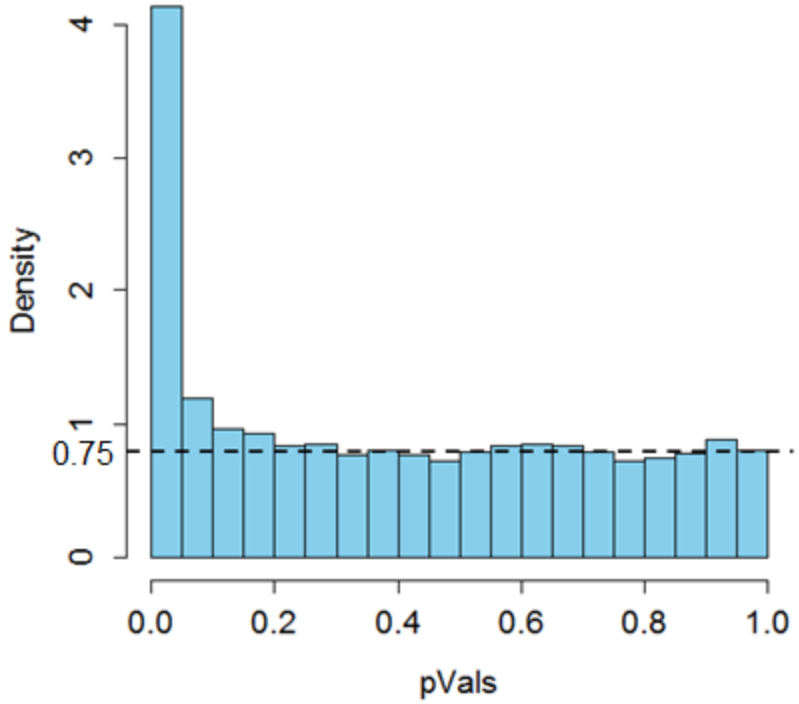
A biologist has developed a cell based assay to investigate the effect of a certain compound on the gene expression of human blood cells. He extracted blood cells from 20 randomly chosen subjects and treated 10 blood samplest with the compound solved in DMSO and the other 10 blood samples only with DMSO. Then he performed a whole genome expression analysis and tested for each of the ~20'000 known human genes if it is significantly differently expressed in compound-treated blood cells compared to the solution-treated blood cells.
For each of the 20'000 tests he got one p-value. The distribution of the p-values is shown in the histogram below.

Use the histogram to roughly estimate the following quantities:

a) How many of the 20'000 genes were unchanged by the compound treatment. Briefly explain your procedure.
The height of the horizontal asymptote gives the proportion of tests with true H_0 corresponding to no compound effect. Here, the height of the horizontal is 0.75 meaning 75% of the 20'000 genes = 15'000 genes did not change their expression upon treatment with the compound.

b) How many significant tests did he get when working with a significance level α=0.05?
The area of the scaled histogram is 1. The area of each bar corresponds to the proportion of observations falling in the respective bin. The bin of the first bar corresponds to p-values < 0.05. The area of the first bar is roughly 0.05*4=0.2 meaning that 20% of the 20'000 test = 4000 tests yielded a significant result -> #PostiveResults=4000.

c) How large is the false discovery rate FDR (=proportion of false positives among significant test results)?
The area in the first bar which is below the horizontal asymptote gives the proportion of tests which gave a false positive result. Here this area can be estimated by 0.05*0.75=0.0375 meaning 3.75% of all 20'000 tests =750 tests yielded a false positive result -> #FalsePositives=750. The FDR is given by #FP/#P=750/4000=0.1875 meaning we have a false discovery rate of 18.75%.


Density
0
1
2
3
4
0.75
0.0
0.2
0.4
0.6
0.8
1.0
pVals