BIN504 - Prob. & Stat. Modeling
for Bioinformatics **Homework II** Fall 2023

> Upload your solutions (as a zip file) to the METU Online site for the course. The **deadline is Jan. 7th at 23.59 + 1 min**. Late submissions will incur a -1% penalty per hour after the deadline.

In this homework assignment you will analyze a set of gene expression experiments conducted on oral mucosa of smokers and nonsmokers. A total of 80 people gave tissue samples with the following distribution:

	Smoker	Non-Smoker
Male	20	21
Female	19	20

Each sample was measured for gene expression on an Affymetrix Human Genome U133 Plus 2.0 microarray chip. So, 80 chips were used. The U133 consists of 54,675 probes which represent more than 38,500 genes. The results of the gene expression measurements are given as four files (one for each group) in the `data` directory in the `HW2.zip` file.

The files are comma separated text files (CSV) with each line being the expression level for a given probe on the U133 chip. The first column is the ID of the probe, the remaining columns are the expression levels of each sample in the group. Samples are labelled with the patient number and type. For example, `NS_F.P101` is patient 101 who was a female non-smoker.

Theoretically there is some amount of variability (due to manufacture) from one chip to the other, but Affy chips are much less variable than traditional microarrays and furthermore the data has been normalized using RMA[1]. Therefore, you may skip any normalization and neglect any technical variability between chips. In other words, you may assume the difference between expression levels for any probe is due to actual biological differences.

Given the information above:

1. (40 points) Which test would be appropriate in order to find differentially expressed probes across smokers and non-smokers? Decide on a test and write the R code that will calculate the raw p values for each probe:
   (a) Using only males (i.e. male smokers vs. male non-smokers)
   (b) Using only females (i.e. female smokers vs. female non-smokers)
   (c) Using combined groups (i.e. all smokers vs. all non-smokers)
   (d) Using ANOVA (i.e. 4 groups separately)
   (e) At $\alpha = 0.05$, find the number of probes that are significantly differentially expressed in each case. Are the findings different? If yes, how significant is the difference[2]?

[1] Robust Multichip Average, see: Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. (2003), *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance.* Bioinformatics 19(2):185-193

[2] *Hint:* You can use a non-parametric test like χ^2 for this

2. (20 points) Use multiple test correction to adjust p values from Question 1, part (c) for $FDR = 0.05$.

   (a) Write R code that will take your results from question 1(c) to incorporate multiple test correction (using `mt.rawp2adjp` in the `multtest` package)

   (b) What are the significantly differentially expressed probes after Bonferroni correction?

   (c) What are the significantly differentially expressed probes after Benjamini-Hochberg correction?

3. (20 points) Using BioConductor in R, you can download the gene annotations for the probes (i.e. which probe ID corresponds to which gene).

   To install the U133 Annotation DB:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite()
> biocLite("hgu133plus2.db")
```

   To lookup the gene name for a given probe (e.g. '1405_i_at'):

```
> library(hgu133plus2.db)
> x <- hgu133plus2SYMBOL
> mapped_probes <- mappedkeys(x)
> xx <- as.list(x[mapped_probes])
> xx["1405_i_at"]
$`1405_i_at`
[1] "CCL5"
```

   Look up the gene names for the differentially expressed probes from Question 2. Compare your findings with the paper the data comes from:

   Boyle JO, Gümüş ZH, Kacker A, Choksi VL et al. *Effects of Cigarette Smoke on the Human Oral Mucosal Transcriptome.* Cancer Prev Res (Phila) 2010 Mar; 3(3):266-78. PMID: 20179299

   Do the genes reported in the paper match the genes you found?