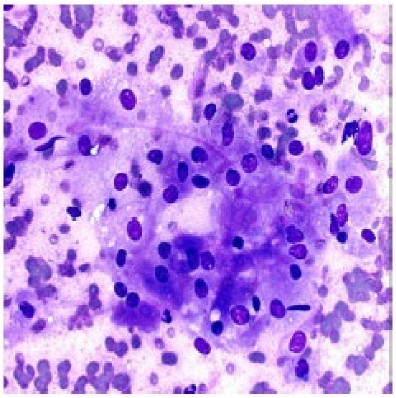
# Instructions for the Group Exercise on Covariance and Correlations

We will be working with data from Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Cancer Letters, 77(2-3), 163-171. Back in 1994 (sic!) this paper used machine learning to predict whether a breast tissue biopsy was cancerous or benign. Here is an example of the image data they worked with:



Black dots are cell nuclei. Irregular shapes or highly variable cell sizes can mean cancer, but it's tricky.
The sample contains 212 cancer patients and 357 healthy individuals (variable *cancer_yn*).
Columns 1 through 30 of the table *cancerwdbc* contain 30 other aggregate characteristics of each patient's biopsy:

| | |
|---|---|
| 1 | radius |
| 2 | texture |
| 3 | perimeter |
| 4 | area |
| 5 | smoothness |
| 6 | compactness |
| 7 | concavity |
| 8 | concave points |
| 9 | symmetry |
| 10 | fractal dim |
| 11 | radius std |
| 12 | texture std |
| 13 | perimeter std |
| 14 | area std |
| 15 | smoothness std |
| 16 | compactness std |
| 17 | concavity std |
| 18 | concave points std |
| 19 | symmetry std |
| 20 | fractal dim std |
| 21 | radius extreme |

| | |
|---|---|
| 22 | texture extreme |
| 23 | perimeter extreme |
| 24 | area extreme |
| 25 | smoothness extreme |
| 26 | compactness extreme |
| 27 | concavity extreme |
| 28 | concave points extreme |
| 29 | symmetry extreme |
| 30 | fractal dim extreme |

The names of these features are listed in the *feature_names* variable.

**Assignment 3A (synthetic data):**

- Generate a sample with Stats=100,000 of two Gaussian random variables r1mix and r2 that have a mean of zero and a standard deviation of 2, and are
  - Case 1: Uncorrelated
  - Case 1: Correlated with correlation coefficient 0.9
  - Case 1: Correlated with correlation coefficient -0.5
- Trick: First create uncorrelated variables r1 and r2 using the rand command. Then create a new variable: r1mix=mix.*r2+(1-mix.^2)^0.5.*r1; where mix= correlation coefficient
- For each value of the mix calculate the covariance and the correlation coefficient between r1mix and r2
- In each case make a scatter plot: plot(r1mix,r2,'k.'); How do the scatterplots for cases 1, 2, and 3 differ from each other? Illustrate the differences by including screenshots in your report.

**Assignment 3B (cancer data):**

- Download the file cancer_wdbc.mat and load it into Matlab using > load cancer_wdbc.mat (be sure to save the file in your current Matlab directory)
- Data in the table *cancerwdbc* (569x30). The first 357 patients are healthy. The remaining 569-357=212 patients have cancer. This information is contained in the variable *cancer_yn*
- Calculate and report the correlation coefficients between each of the 30 characteristics and the presence/absence of cancer. Use the Matlab command [cancer_corr, p_corr]=corr(cancer_yn, cancerwdbc); the first output variable (cancer_corr) is the Pearson correlation coefficient, and the second variable (p_corr) is the probability of getting this or a stronger correlation by pure chance. One can only report correlations for which this probability is below 0.05. List all features, correlations, and p_corr probabilities that meet this criterion.
- Generate scatter plots of Mean Area vs. Mean Perimeter and Mean Texture vs. Mean Radius (see above for which columns correspond to these variables).
- Compute and report the Pearson and Spearman correlations of these two pairs of variables.
- Compute the matrix of Pearson correlations between all pairs of variables: there are 30*29/2=435 such pairs. Hint: corr_mat=corr(cancerwdbc); does the trick of calculating the matrix of all correlations.
- Plot a histogram of these 435 correlation coefficients. Include the screenshot of this histogram in your report.