**Practical 4**

**Variance, Covariance and Correlation**

Open the Excel spreadsheet “Exercise Sheet GeneDataAnalysis 1”.

Exercise Sheet GeneDataAnalysis 1

1. In box 1 compute sample size (N), degrees of freedom (df), sum, mean, variance and standard deviation for feature A.

The functions to use are:

N =COUNT(C2:C11)

df =C13-1

sum =SUM(C2:C11)

mean =AVERAGE(C2:C11)

variance =VAR(C2:C11)

stdev =STDEV(C2:C11)

Copy the cells containing the functions to the right in order to compute the descriptive statistics for all four features.

2. Go to TABLE 2.

In the first cell of the table (I2) write the command: =C2-C**$**16

This subtracts the mean of feature A (computed in cell C16) from the first measurement outcome of feature A (which is stored in cell C2).

Mind the dollar sign! What is its function?

Copy the cell with the command down to cell I11 in order to subtract the mean of A from each of its measurement outcomes. This *scaling procedure* is called “centring”.

Select the computed range and copy it to the right in order to compute the centred values for all four features.

3. Select the descriptive statistics computed in BOX 1 (i.e. the cell range A13:F18). Copy and

Paste the selected range into BOX 2 (in cells I13:L18). This gives the descriptive statistics

Of the centred features.

Notice what has happened to the mean, var and stdev.

4. Go to BOX 3.

In the first cell of the box (I20) write the function: =SUMSQ(I2:I11). What is it doing?

In the cell below cell I20 divide SS by df by typing in the command: =I20/I14.

Select the two cells I20:I21 and drag the selcted range to right in order to copy the commands so they can repeat the computation for the other features.

Look at the results and compare them with the descriptive statistics above. What do you notice?

5. Go to TABLE 3.

In the first cell of the table (O2) write the command: =**$**I2\*J2 and copy it down.

Select the just computed range and copy it to the right in order to apply the same computation to the other features.

What is the purpose of the dollar sign ($)?

Select the first three lines and columns of BOX 2 (I13:K15), copy and past them into BOX 4.

6. Go to BOX 5.

In the first cell of the box (O20) type the function: =SUMPRODUCT(**$**I2:**$**I11,J2:J11).

What is this function doing?

Divide the SP by the degrees of freedom and copy the two cells to the right (as done in Step 4 for BOX 3).

Mind the **dollar signs**!

In cell O22 (the last cell of the first column of BOX 5) type the function: =COVARIANCE.S(**$**I2:**$**I11,J2:J11)

and copy the cell two places to the right.

Check the values. Conclusion?

7. Go to TABLE 4.

In the first cell of the table (T2) write the command: =I2/I**$**18 and copy it down.

What is the command doing?

The *scaling procedure* performed by the command is called **standardisation** and the resulting values are called **Z-scores**.

Select the values of the first column of TABLE 4 just created and copy them to the right in order to standardise the values of the other features.

8. Complete Boxes 6 – 9 and TABLE 5 in the same way as done for boxes 2 – 4 (Steps 3 – 5) and TABLE 3 (Step 5).

In cell Z23 type the function: =CORREL(**$**C2:**$**C11,D2:D11) and copy it two cells to the right.

Check the **cell ranges** this function is working on and compare the outcomes with those of the row above it.

Conclusion?

9. Open the spread sheet “Exercise Sheet GeneDataAnalysis 2” next to “Exercise Sheet GeneDataAnalysis 1”.

Copy the values of feature “A” from TABLE 1, “Exercise Sheet GeneDataAnalysis 1” into the input box (under the header “x”) of “Exercise Sheet GeneDataAnalysis 2”.

Copy the values of feature “C” from TABLE 1, “Exercise Sheet GeneDataAnalysis 1” into the input box (under the header “y”) of “Exercise Sheet GeneDataAnalysis 2”. Do not tamper with the “1” scores in the “on/off” column!

Follow Instruction 3 on the spread sheet.

Compare the outcomes of the computations with the ones you did above in the first spread sheet and have a look at the scatter plot.

Repeat this exercise by copy and pasting the values of feature “G” into column “y” of the input box, but do not change the x values.

Consult the slides of Lecture 5 for an explanation of this spread sheet.