DEPARTMENT OF PHYSICS & ASTRONOMY
3459 EXAM-2
14:00 - 17:00 : 12th December 2011

Please read the exam guidelines, rules, instructions and marking criteria at
`http://moodle.ucl.ac.uk/mod/wiki/view.php?id=13963&page=Final+exam`
(linked from the *Exams and Coursework* page).

This exam is worth 50% of your final mark for the course. The duration of the exam is 3 hours. Students should upload the Java source code files for their solution using Moodle under the section headed "Exam 2".

In this exam you will be processing data on the occurrence of several diseases within the United Kingdom, broken down into geographical regions. You will read the data provided in a web directory, store them into (a) suitable container(s), and perform various statistical analyses.

Four files are provided in the following web directory:
http://www.hep.ucl.ac.uk/undergrad/3459/exam-data/2011-12/.

- The file regions.txt contains a list of UK regions: the first field on each line is an identification ID string, and the rest of the line is the name of the corresponding region. The region names can contain spaces; the ID string and region name are therefore separated by a comma.
- The file populations.txt contains the total population of all regions. The first field on each line is the ID string of the region, and the second field is its population.
- The file occurrencesXYZ.txt gives the number of disease occurrences in one year. Each line contains the following fields:
  - the identification ID string of the region;
  - the number of XXX cases in that region;
  - the number of YYY cases in that region;
  - the number of ZZZ cases in that region.
- The file occurrencesAB.txt gives the number of disease occurrences in one year. (NB: due to unavailability of data some regions do not have entries in this file.) Each line contains the following fields:
  - the identification ID string of the region;
  - the number of AAA cases in that region;
  - the number of BBB cases in that region;

Part 1: 15/50 marks

Write a program to do the following:

- Read all the data from regions.txt, populations.txt, occurrencesAB.txt and occurrencesXYZ.txt, and store them in an(some) appropriately designed data structure(s).
- Summing over all regions, find the total UK population under consideration and the total number of occurrences per capita (per person) for each disease, and print out these values. (Note: since some of the regions did not report data for all of the diseases, the population under consideration is not the same for all of the diseases.)
- Find the region with the highest and the region with the lowest number of total occurrences of all disease per capita, and print their names.

Part 2: 20/50 marks Make the following enhancements to your code:

- Define an interface to calculate any statistic (e.g. mean, median, maximum, RMS, covariance, etc.) from the data for a particular disease or pair of diseases.
- Use this interface to calculate the following:
  - For each disease, find the region with the highest statistical excess of occurrences. Print the name of the region, the number of occurrences and the statistical significance. For a Poisson distributed variable the statistical significance can be approximated by:

$$S_{xi} = \frac{x_i - p_i}{\sqrt{p_i}} \tag{1}$$

    where x_i is the measured number of occurrences of disease x in region i and p_i is the expected number of occurrences based on scaling the national per capita rate to the population of region i.
  - Find the correlation of the occurrences of disease XXX with diseases YYY, ZZZ, AAA and BBB. The correlation of two variables is defined as:

$$\phi_{XY} = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N\sigma_x\sigma_y} \tag{2}$$

    where i represents each of the N regions, $\bar{x}$ & σ_x and $\bar{y}$ & σ_y are the mean and standard deviation of x and y.

Part 3: 15/50 marks Make the following enhancements to your code:

- Define an interface that filters out the regions that satisfy any given condition based on a statistics interface (as defined in part 2). The filtering interface method should return the same structure as the original disease data, but should only contain regions satisfying the provided condition. Implement the interface with the appropriate class holding the data.
- Use this interface to create a database containing only the 5 regions with the highest and the 5 regions with the lowest number of occurrences per capita of disease XXX. Print a list of these regions to the screen, sorted according to the number of occurrences per capita.

Uploading your work

If you use your own classes from earlier modules, make sure you upload them as well as any new classes you create during the exam.

END OF PAPER