CS 237 ASSIGNMENT 1 – Part 2

Michalis Faloutsos

Please read carefully.

General notes. Work in pairs, but work together, as opposed to splitting tasks. Obviously, do not copy from other pairs, though you could discuss.

Explain your answers with brief but to the point comments. Simple numerical answers are not enough. The assignments should be readable; one should jsut read you assignments and understand what is happening even if they don't know the questions.

Warning: I may choose pairs to explain their work to me. Both persons should be familiar with all of the assignment to the extent of explaining me how things work.

IMPORTANT: For the programming questions.

A. The programs should be written in script language: awk, shell, perl. The use of awk is strongly suggested. You can also find a sample awk file and script in my web-page. Java, C, C++ etc are not acceptable; the point is to learn how to use scripts for which you will be thanking me for the rest of your lives.

B. You should hand-in input, output files and the code.

1. Assume the measurements of queries per job as follows: [35, 28, 15, 13, 38, 24, 34, 36, 21, 39, 34, 23, 45, 33, 28, 42, 23, 33, 14, 15].

   (a) What is the 10-percentile and 90-percentile of the samples?

   (b) What is the average?

   (c) What is the median?

   (d) What are the modes?

   Note: Show the formulas that you use and some intermediate results. One number answers will get only partial credit. You may use the scripts in the next questions to test or calculate your answers.

2. Write a small program to test for zero mean. Assume you compare two systems and the data is given in pairs in a file of two numeric columns. Consider the paired observation case.

3. Write a small program to determine the sample size. The input is a small sample file with one column and the accuracy r. The output is the number of samples N we need to achieve a 90% confidence interval with an accuracy of r% (the real population mean should be at most r% from the mean that we will calcuate using the N samples).

4. Calculate the distributions.

   (a) Consider the independent variables x $N(\mu, 1)$ and y $N(\mu, 1)$.

      i. $E(x), E(y)$

    ii. $E(x) - E(y)$
    iii. $E(x) + E(y)$
    iv. $(E(x) + E(y))/2$

(b) Consider two samples $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_2\}$. Calculate the distributions.

    i. $\bar{x}, \bar{y}$
    ii. $\bar{x} - \bar{y}$
    iii. $\bar{x} + \bar{y}$
    iv. $(\bar{x} + \bar{y})/2$
    v. $(\bar{x} - \mu)/\sqrt{(n)}$

5. Linear Regression. Write a small program to calculate the linear regression parameters of a two parameter model $(x, y) : y = b_0 + b_1 x$. The input is a) a number of x_i, y_i measurements $i = 1, ..., N$ in a two column file. Note that N is not known, until the file is read. of The output should be: a) the regression parameters b_0, b_1, b) sample correlation or Pearson's r, c) the 90% confidence intervals for the regression parameters.

6. Curvilinear regression. Write a small program to calculate the regression parameters of an exponential model: $(x, y) : y = bx^a$. Input and outputs are the same as in the Linear Regression question. Hint: You should be able to transform the model to a linear model and reuse the code from the previous question. In other words, transform the data to a linear model, solve it, and then transform the results.

That will be all for now.