**T-122.103 EXERCISE 2/2003**

In T4 on 26 September 2003 at 12:30–14 o'clock. The next exercise sessions beginning from 3 October 2003 will begin at 12:15 o'clock.

Recall the notation for sets of variables: $ABC = \{A, B, C\}$ etc.

1. Study the data set "smallilmo" given in the course web site. Find out the number of attributes (variables), observations, distribution of number of 1s per observation, most highly correlated pairs of variables, etc.

2. Consider a 0-1 dataset $D$ over the attributes $A$, $B$, and $C$. Suppose we know that $f(A) = 0.4$, $f(B) = 0.3$, $f(C) = 0.5$ and $f(AB) = 0.29$. What can be said about the frequencies of $AC$, $BC$ and $ABC$?

3. Suppose $ABC$, $BCD$, $ABD$, $ABE$, $BCE$, $ACD$, $ADE$, and $BDE$ are frequent. Which sets of size 4 can be frequent?

4. Consider a data set with $n$ observations and $m$ variables. Assume each entry in the data is independent of the others and 1 with probability $p$ and 0 with probability $1 - p$.

   (a) What is the distribution of the frequency of a set $X$ of variables?

   (b) Approximate the probability that a set $X$ of variables has frequency higher than $\sigma$.

   (c) Estimate how many frequent sets of size $k$ there will be in data set.

5. Design a very simple program (e.g., in Matlab) that computes the frequencies of all subsets of a data set. Study the distribution of the frequencies of sets in a random data set.