Final exam

N. Saunier

December 10, 2014

Please

- note the scale (the total score is out of 20) and the indicative time to devote to each exercise;
- clearly indicate the numbers of the questions you are dealing with and your corresponding answers (and underline or frame the numerical results);
- pay special attention to the wording and definition of the notations you use.

You have access to the moodle site for the course, with the course notes and the necessary statistical tables.

The data required for certain exercises are available in the data archive on moodle (Section “ Final exam ”, first element “ Data ”). Text files are provided in a period and comma version for decimal numbers. A spreadsheet (Excel) can be placed on moodle for questions involving calculations and figures from the data provided for certain exercises.

Exercise 1 15 min (/1 pt)

Describe two advantages and two disadvantages of the Origin-Destination surveys carried out in the greater Montreal area every five years.

Exercise 2 25 min (/3 pts)

Present and describe using statistics and figure (s) all the average speed data (in km / h) between two points on a motorway contained in the file `vitesses-bt-b-a.txt`.

Exercise 3 50 min (/8 pts)

This exercise is based on a set of traffic data recorded by magnetic loops in the Portland area, imported into the SQLite `freeway_loopdata1hr.sqlite` file. This file contains the metering and speed data aggregated over one hour periods for several metering stations. The useful columns of the “ loopdata ” table are as follows:

- “ detectorid ”: identifier of the counting station;
- “ starttime ”: start (day, hour and time zone) of the one hour interval on which the traffic data is aggregated;
- “ volume ”: hourly flow (number of vehicles per hour);
- “ speed ”: average speed over the hour (in miles per hour);
- “ occupancy ”: occupancy rate (percentage of the time during which the sensor is occupied by a vehicle);

- “ date ”: date corresponding to “ starttime ”;
- “ time ”: hour corresponding to “ starttime ”;
- “ daytype ”: day of the week (whole number: 0 corresponds to Sunday, 1 to Monday, ... and 6 to Saturday).

Please answer the following questions:

1. What is the primary key of the “ loopdata ” table? Does the “ loopdata ” table follow the three normal forms? Justify your answer. (1 pt)
2. For the 1732 counting station, write the SQL query to calculate the average speed and the number of speed measurements per day of the week (Monday, Tuesday, etc.). Perform the appropriate statistical test to determine if the day of the week affects the average speed at that station. (3 pts)
3. Write the query allowing to calculate the average hourly flow per hour for each hour of the day on all the days of the week (Monday to Friday included) for each station. (0.5 pt)
4. Four hourly periods of the day (night: midnight to 6 a.m.; morning: 6 a.m. to noon; afternoon: noon to 6 p.m.; evening: 6 p.m. to midnight).
   (a) Give one of the queries to create one of the four new tables (or views) calculating for each station the average flow per period (one table / view per period) for the days of the week. (0.5 pt)
   (b) Write the query to join the four tables / views to get the average throughput per period of the day per station. (0.5 pt)
   The result looks something like:

Station	Night flow	Morning flow	Afternoon flow	Evening flow
1345	123	456	789	123
1346	456	789	123	123
...				

   (c) Each station is now characterized by four average flows per period of the day: apply the k-averages algorithm (after exporting the table to the `csv` file) to identify homogeneous groups of stations with similar flows according to time of day. Present the results in a few lines. Represent the centroids of the groups on a figure. (2.5 pts)

Exercise 4 30 min (/4.5 pts)

This exercise is based on the car characteristics dataset contained in the `autos.txt` file. It aims to study the relationship between the two variables length and width of cars (columns “ length ” and “ width ”). Please answer the following questions:

1. Plot the scatter plot of width versus length and calculate the correlation coefficient: comment. (1 pt)
2. Using one of the software at your disposal, estimate the linear regression line of the width as a function of the length:

(a) Discuss the significance of the model and calculate (without using Excel again) the confidence interval at 90 % and 95 % of the coefficient a of the length noting that the statistic $\frac{\hat{a}-a}{s_{\hat{a}}}$ follows a Student law with $n-2$ degrees of freedom (with $s_{\hat{a}} = \sqrt{\frac{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2}{\sum_i (x_i - \bar{x})^2}}$, y_i and x_i respectively the width and length of the vehicle i, $\bar{x}$ the empirical mean length and ˆ denoting the estimated or predicted terms). (2.5 pts)

(b) Based on the graphical study of the residuals, indicate whether the linear regression assumptions are met. (1 pt)

Exercise 5 30 min (/3.5 pts)

A trajectory is a set of measurements of positions (x, y) at instants t.

1. Propose a relational data model allowing to store trajectories and to make queries on their positions. Clearly indicate the primary key and types of attributes. (1 pt)
2. We wish to carry out a pilot project of the movements of a small group of drivers by GPS receiver. Modify the previous database to record characteristics of each driver participating in the project and their GPS trajectories (eg name, age, sex, place of residence, etc.). Always indicate the primary key and the types of the attributes. (1 pt)
3. Write an SQL query for the extraction of the movements of the user Paul, by ordering the positions of his trajectories temporally. (1 pt)
4. What feature of database management systems protects the individual information of users participating in the project? (0.5 pt)