Midterm Exam

N. Saunier

October 20, 2021

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 particular attention to the wording and definition of the notations you use;
- note that some exercises require files available on Moodle (" Midterm Exam " section) (the text files are provided in a version with a period and a comma for decimal numbers, if necessary). Statistical tables are available on Moodle if necessary.

Exercise 1 (data collection) 45 min (/7.5 pts)

1. Name and describe two constraints that may limit the process of acquiring or collecting data. (1 pt)
2. Two hypothetical situations are presented below. For each situation, please (2 pts)
   - determine the characteristic of the data collection method that will be problematic when processing and analyzing the data;
   - Determine if the problem is related to the variability of the observations, to a bias in the data or to the ease of processing the data.

   (a) An on-board survey was carried out in order to draw up a portrait of the mobility of all users of the suburban trains of the Metropolitan Transport Network (RTM, exo). To this end, a questionnaire was rigorously constructed, and the interviewers received training to avoid influencing the responses of the respondents. For survey purposes, a large sample is desired in order to increase precision. Exo therefore decides to survey all users of the train line to Mont-Saint-Hilaire, one of the busiest lines on the network, during a typical weekday in November. The questionnaires are filled in by the interviewers, directly on a tablet from the users' responses.

   (b) A major boulevard is congested during rush hour. During this period, a significant number of trips seem to be reallocated on a local and residential street parallel to the boulevard. After several complaints from residents, you are mandated to characterize the transit in the area and to assess the transit rate at peak periods. You therefore call on a supplier of data from

cell phones (geolocated travel data) to assess the origin and destination of all the trips made on the local street.
The sample is estimated at 10 % of all vehicles passing through the local street and data is provided for each day of the week as an average of the whole of 2020. Data from the provider have the following format:

Id	Origin	Destination	Day	Number of observations
1	Street A	Street B	Monday	54
2	Rue A	Rue B	Tuesday	48
...				
326	Rue C	Rue D	Monday	60
...				

3. You take a speed survey by radar to determine the average speed on a road during rush hour. What is the minimum sample size you need to collect to determine the mean speed on the axis if you want a 95 % confidence level and absolutely want to keep a margin of error (tolerance) of 3 km/h maximum? You know that the flow of the road studied at rush hour is 1300 veh/h. Several similar studies previously carried out on similar axes show that the standard deviation of speeds varies from one road to another between 5 and 14 km/h. (1.5 pt)

4. For the same route as the previous question, we also want to know the proportion of trucks: what is the size of the sample necessary to determine it with a margin of error (tolerance of 4 %) with a level of 95 % confidence? What is the sample size required to record speeds and vehicle type in a single collection? (1.5 pts)

5. Give the definition of the reference population and the sampling frame? Give an example where they are the same. (1.5 pt)

Solution

1. The course mentioned five constraints for data collection:
   - cost
   - privacy
   - availability of a collection method
   - data property
   - storage

2. Here are some answers to the questions:
   - The spatial coverage is deficient (survey on a single line of the RTM, while the survey targets all users) and will lead to a bias in the results.
   - The temporal resolution of the data used from cellphones is deficient (the study targets peak hours, but the data provides data for an entire year without specifying the hours of travel) and will lead to a bias in the analyzes which cannot be be corrected by data processing.

3. The minimum sample size is $n = \frac{k^2_{\alpha/2}\sigma^2}{e^2} = \frac{1.96^2 14^2}{3^2} = 83.7 \approx 84$ (taking the worst case for the standard deviation of the speeds).

4. The minimum sample size is $n = \frac{k^2_{\alpha/2} p(1-p)}{e^2} = \frac{1.96^2 0.5(1-0.5)}{0.04^2} = 600.2 \approx 601$ (taking the worst case 0.5). One therefor just has to collect the speed and type information (truck or not) for 601 vehicles to have the desired precision on the averages of the two attributes (speed and type of vehicle).

5. The reference population is the set for which we seek to obtain information. A sampling frame is the population used to draw a sample (subset) that will be surveyed, and which may be different from the target population depending on the data sources available. The sampling frame is the reference population, for example when the subscribers of a service are surveyed and the database of subscribers is used as a sampling frame.

Exercise 2 (database and SQL) 30 min (/4 pts)

Download the database `ReseauRoutier.db` of the midterm exam section on Moodle which describes a road network, containing three tables for intersections (“ nodes ” table), road segments (“ sections ” table) and the turns allowed at crossroads (“ turnings” table). Indicate the SQL queries which allow the following questions to be answered (it is not required to give the result of the query). The “ DB Browser for SQLite ” software is available on computers if you wish to test your queries.

1. Calculate the average speed limit (“ speed ” field) on the sections of the whole network, without weighting, then weighted by the length (“ length ” field) of the sections. (1 pt)

2. Calculate and display the number of sections for each class of speed limit. (1 pt)

3. Display the whole of the turning movement table and add the attribute that defines the type of junction (“ nodetype ” field of the “ node ” table) in order to specify the type of intersection in which the turning movement is located. (1 pt)

4. Decreasingly display the attributes of intersections according to the number of turning movements subject to a stop sign (knowing that a movement managed by a stop is represented by the value 2 of the “ sign ” field). The list must include the unique identifier of the intersection, the name of the intersection (“ name ” field) and the number of movements managed by a stop sign. (1 pt)

Solution

1.
```
SELECT AVG (speed), SUM (speed * length) / SUM (length) FROM sections
```

2.
```
SELECT speed, count (*) FROM sections GROUP BY speed
```

3.
```
SELECT turnings.*, nodes.nodetype FROM turnings, nodes
WHERE turnings.id_node = nodes.id
```

4.
```
SELECT id_node, name, COUNT (turnings.id) as nturnings
FROM turnings, nodes
WHERE turnings.id_node = nodes.id AND turnings.sign = 2
GROUP BY id_node
ORDER BY nturnings DESC
```

Exercise 3 (data model) (/6.5 pts)

We want to design a data model for the travel survey carried out by the City of Montreal using the MTL journey mobile application. After installing the application on their phone, the participants answer a first questionnaire on their socio-demographic characteristics and their transport habits. The application then records their movements (GNSS sensor) for 30 days. For each trip, when it detects the end of the trip, the application asks the respondent for additional information such as the mode and reason for the trip.

1. Propose a model for the data collected with the MTL trip mobile application in the form of an Entity / Association diagram involving at least the following entities: participant, trip, GNSS point. Add attributes (including the identifier) and associations between entities, with their minimum and maximum cardinalities, and functionalities. (1.5 pt)
2. Translate the Entity / Association schema into a relational schema. Clearly indicate primary and external keys, and suggest types for attributes. (1 pt)
3. Discuss what kind of spatial data (matrix or vector) the GNSS data corresponds to, with two advantages and two disadvantages of this type of data. Give an example of data of the other type relevant to transport. (2 pts)
4. Propose a coordinate reference system adapted to the data collected by the MTL Trajet application and justify the choice. (1 pt)
5. To take advantage of the functionalities of a geographic information system (GIS), we want to use a spatial database to record movements. Discuss how to modify the data model using spatial data types to allow direct display of user movements in a GIS. (1 pt)

Solution

1. The entities and their attributes are as follows (the identifier of each entity is in **bold**):
   - **participant id**, name, date of birth, occupation, postal code of residence
   - **trip id**, mode, purpose
   - **GNSS point id**, date, time, (x,y) coordinates

   The associations are as follows (it is desirable to name the associations and to make a diagram):
   - participant-trip: a participant makes 0-n trips, one trip ment is done by 1-1 participant. The functionality is 1-n.
   - GNSS point-trip: a trip consists of 1-n GNSS points (we could probably say 2 to n), a GNSS point is part of 1-1 shift. The functionality is 1-n.
2. Each entity becomes a table (Participants, Trips, GNSS Points). It is not no need to add tables to represent n-m associations. It is necessary add the following foreign keys for 1-n associations:

- participantId in Trips referring to Participants.id;
- displacementId in GNSS Points referring to Trips.id.

Here are some data types for attributes: name, occupation, zip code, mode and pattern are categorical attributes, represented by text. The date of birth is of type date (as the date of the GNSS Points table). Contact details x and y are decimals, or text according to the spatial reference system used.

3. GNSS sensors measure the successive positions of users in the form of series of points. Points are vector data. The advantages and disadvantages are discussed in the course notes. An example of data matrix relevant for transport is the digital model data of surface, used for the design of roads and slopes in the calculations of paths for example for cycling.
4. We could use the MTM or UTM coordinate systems (more precise in longitude), where the Montreal area is 18 and 8, respectively.
5. The simplest would be to replace the x and y coordinate fields in the table of points by a point field. An alternative would be to remove the GNSS points table and simply have a row type field in the table Shift. One challenge would be to represent time, without a good solution if the points are not recorded at regular time interval.

Exercise 4 (data processing) 30 min (/2 pts)

1. We have data of floating vehicles moving on the network for collecting travel times on journeys and we record their successive longitudinal positions (distance traveled as a function of time) at regular time intervals. Indicate the output of the following algorithm (which is calculated): (1 pt)

```
entry: $n$ longitudinal positions $d_1, ..., d_n$, duration $\Delta t$ between records
position
output:?

start
  $x = 0$
  for $i = 2...n$
    $x = x + \frac{d_i - d_{i-1}}{\Delta t}$
  return $\frac{x}{n-1}$
end
```

2. Modify the previous algorithm so that it measures the length of time the vehicle spends at a standstill. (1 pt)

Solution

1. Here is a possible solution:

- Inputs: a series (list) of n positions (a position is a point, which is a list of dimension 2), transport networks (one graph per transport network), a point map of interest (workplaces, shops, restaurants, cafe, etc.)
- Output: a list of size $n-1$ of the mode used to move between each position

2. The algorithm presented calculates the average speed of the movement.
3. Here is a solution (note that we could add a small tolerance for the non-displacement test (d_i equal to d_{i-1}), such as a maximum distance threshold related to the noise of the GNSS positions in absence of movement):

```
**entry**: $n$ longitudinal positions $d_1, ..., d_n$, no time $\Delta t$
**output**:?

**start**
  $dure_{stop} = 0$
  **for** $i = 2...n$
    **si** $d_i$ equal to $d_{i-1}$
      $dura_{stop} = dura_{stop} + \Delta t$
  **return** $dura_{stop}$
**end**
```