Midterm Exam

N. Saunier - F. Bélisle

Octobre 3rd, 2023

Please

- note the score (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.

Exercice 1 (data collection) 45 min (/6 pts)

1. Name and describe two constraints that may limit the process of acquiring or collecting data. (1 pt)
2. Avenue du Mont-Royal, the busiest commercial street in Montreal, is closed to automobile traffic for part of the summer. An experiment has been carried out since the summer of 2021 with the authorization for cyclists to remain on their bikes on the pedestrian street. Pedestrians complain about bicycle traffic. In this context, please propose and describe different data collection methods for the following purposes (if recommending a survey, please specify the target population, survey type and method (medium) and time frame): (3 pts)
   (a) to measure bicycle traffic and their speed;
   (b) to study the reasons for pedestrian complaints;
   (c) to determine whether incidents occur on the pedestrian street.
3. Calculate the sample size needed to determine the proportion of cyclists with a tolerance of 4 % for a confidence level of 95 %. Should more data be collected to determine the average speed of cyclists with a confidence level of 95 % and a tolerance of 1 km/h (knowing that the standard deviation of the speed of cyclists was measured at 6 km/h on the St-Denis Street cycle path. (2 pts)

Solution

1. The course mentioned five constraints for data collection:
   - cost

- respect for private life
- availability of a collection method
- data property
- storage

2. (a) Observation methods for collecting point data are suitable. Temporary methods (albeit for several days or weeks) are also preferable. Finally, it is necessary to be able to classify users, at least distinguish pedestrians from cyclists. Pneumatic tubes (in pairs for speed) seem most suitable, as well as video cameras (to have better classification performance).

   (b) The only way to know the motivations for people's behavior is to conduct a survey (a poll) among users. This would be a sectional survey (for a limited period) by interception (this is also possible face to face, but takes more time since the investigator asks the questions and notes the answers). The target population is made up of all pedestrians who use Avenue du Mont-Royal. The time frame should cover the entire day and evening to avoid excluding pedestrians who travel at specific times (such as the evening or early morning).

   (c) Determining whether incidents are occurring requires a spatial data collection method. Incidents can occur at different locations on the street and take place in one area and over a certain time, which cannot be captured at one point on the street. Video cameras (or LIDAR) are therefore appropriate, depending on their ability to detect and classify incidents. It will also be possible to manually validate the data, and check if other incidents (than those detected automatically) occur. Direct manual observations are also possible, although more expensive, which will reduce the observation period. Finally, user surveys would be possible, but will suffer from the subjectivity of people and their ability to remember the details of incidents.

3. The necessary number of user observations to determine the proportion of cyclists with a tolerance of 4 % for a confidence level of 95 % is $z^2_{0.975}\frac{\hat{p}(1-\hat{p})}{e^2} = 1.96^2\frac{0.5(1-0.5)}{0.04^2} = 600.25 \approx 601$ (the proportion of cyclists not being known, we consider the worst case $\hat{p} = 0.5$). To determine the average speed of cyclists with a confidence level of 95 % and a tolerance of 1 km/h (with the empirical standard deviation $s = 6$ km/h), the number of observations number of cyclists is $z^2_{0.975}\frac{s^2}{e^2} = 1.96^2\frac{6^2}{1^2} = 138.30 \approx 139$. There would therefore be enough observations to determine the proportion of cyclists, i.e. 601. This was not asked, but it actually depends on the proportion of cyclists. In fact, there must be at least 139 cyclists among the 601 users observed to guarantee accuracy on the average speed of cyclists.

Exercice 2 (Databases and SQL) 30 min (/6 pts)

Québec's Pedestrian Association seeks to address the issues of sidewalk quality (defects, coating) and associated accessibility. It is therefore necessary to create a database so that volunteers can record their observations.

1. Propose a data model in the form of an Entity/Association diagram that allows you to record information about sidewalks. The model will consider at least the following entities: sidewalk, street, crossing (for pedestrians), street furniture

(like lamppost, bench, bollard), sidewalk problem (for walking), and bus stop. Add attributes (indicating identifiers) and associations between entities with their cardinalities, minimum and maximum, and functionalities. (4 pts)

2. Translate the Entity/Association schema into a relational schema. Clearly indicate primary and external keys (and what the external keys refer to), and provide types for the attributes. (2 pts)

Solution Even if the solution is provided here in the form of lists, it is required to make a *diagram*, a *graphic* representation.

1. The entities and their attributes are as follows (the identifier of each entity is in **bold**):

   sidewalk : **id**, width, length, type of surface

   street : **id**, name, number of lanes, direction of traffic (one or two)

   crossing : **id**, width, type of paint, presence of traffic control

   street furniture : **id**, type, footprint (floor space)

   problem : **id**, type, observation date, resolution date

   bus stop : **id**, bus line, presence of a shelter for pedestrians, presence of a bench

   The associations are as follows (it is desirable to name the associations and draw a diagram):

   - street-sidewalk: a street sometimes has no sidewalk and has up to two (0-2) (generally, but we can discuss streets with central reservations), a sidewalk is part of one and only one street (1-1). The functionality is 1-n (often n=2).
   - sidewalk crossing: a crossing connects two sidewalks (exactly a priori (cardinality 2-2); there is generally no marked pedestrian crossing in the absence of a sidewalk). A sidewalk has zero to several crossings. The functionality is 2-n.
   - street furniture-sidewalk: street furniture (bench, lamp post, trash can) is generally located on a sidewalk (exactly, cardinality 1-1), a sidewalk can include zero to several elements of street furniture. The functionality is 1-n.
   - sidewalk problem: by definition of the subject, we are interested in identifying sidewalk problems. A sidewalk can have zero to n problems, one problem is on exactly one sidewalk (1-1). The functionality is 1-n.
   - bus stop-sidewalk: it is difficult to imagine a bus stop without a sidewalk (it must exist), in which case a sidewalk can have zero to n bus stops, and a bus stop is located on exactly one sidewalk . The functionality is 1-n.

   It should be noted that other choices were possible, such as associating crossings, problems, street furniture and bus stops with the street rather than the sidewalk. We must also distinguish the types of problems (for which we would no longer have to create an entity) from the problems themselves.

2. Each entity becomes a table (Sidewalk, Street, Crossing, Street furniture, Problem, Bus stop). It is not necessary to add tables to represent n-m associations. The following external keys must be added for 1-n associations:

   - streetId in Sidewalk referring to Street.id;

- sidewalkId1 and sidewalkId2 in Traverse referring to Sidewalk.id.
- sidewalkId in Urban furniture referring to Sidewalk.id;
- sidewalkId in Problem referring to Sidewalk.id;
- sidewalkId in Bus stop referring to Sidewalk.id.

Here are some types of data for the attributes: the type of surface, direction of traffic, type (Street furniture and Problem), bus line and the "presence" attributes for Bus stop are categorical attributes, represented by text . "date" attributes are of type date. The length, width and footprint attributes are decimal numbers, number of lanes is an integer.

Exercice 3 (Data Treatment) 30 min (/2 pts)

1. We have data from floating vehicles moving on the network for the collection of travel times on routes 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 (what is calculated): (1 pt)

   input: n longitudinal positions $d_1, ..., d_n$, duration Δt between recordings of positions
   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 driver performs hard braking (high deceleration). (1 pt)

3. Bonus point: modify the algorithm to count distinct hard braking events.

Solution

1. The algorithm presented calculates the average speed of movement.

2. Here is a solution:

   input: n longitudinal positions $d_1, ..., d_n$, time step Δt,
   sudden braking threshold $v_{brake} < 0$
   output: duration of time with sudden braking

   start
   $duration_{braking} = 0$
   $v_2 = \frac{d_2 - d_1}{\Delta t}$
   for $i = 3...n$

$v_i = \frac{d_i - d_{i-1}}{\Delta t}$
si $\frac{v_i - v_{i-1}}{\Delta t} < v_{brake}$
$duration_{braking} = duration_{braking} + \Delta t$
send $duration_{braking}$
end

3. Here is a solution (it involves counting the events: as long as the braking is lower than the threshold value, it is the same event):

input: n longitudinal positions $d_1, ..., d_n$, time step Δt,
sudden braking threshold $v_{brake} < 0$
output: number of sudden brakings

start
$n_{braking} = 0$
$braking = false$
$v_2 = \frac{d_2 - d_1}{\Delta t}$
for $i = 3...n$
$v_i = \frac{d_i - d_{i-1}}{\Delta t}$
si $\frac{v_i - v_{i-1}}{\Delta t} < v_{brake}$
si $braking$ false
$braking = true$
$n_{braking} = n_{braking} + 1$
otherwise if $braking$ true
$braking = false$
send $n_{braking}$
end

Exercice 4 (Database and SQL) 45 min (/6 pts)

This exercise is based on the database provided on parking signs on the road posts `signalisation-stationnement.sqlite` available on Moodle (data coming from the City of Montreal's open data portal). There are two tables. The description of the fields in the `signalisation` table is as follows:

- POTEAU_ID_POT: Identification number of post
- POSITION_POP: Number of the position sign on the post
- PANNEAU_ID_PAN: Identification number of the sign
- PANNEAU_ID_RPA: RPA ID number of sign
- DESCRIPTION_RPA: RPA Description of sign
- CODE_RPA: RPA code of sign
- FLECHE_PAN: Code of the arrow of sign
- TOPONYME_PAN: Toponymic of sign
- DESCRIPTION_CAT: Description of the category of the sign
- POTEAU_VERSION_POT: Version of the sign
- DATE_CONCEPTION_POT: Conception date of the sign
- PAS_SUR_RUE: Indicates wether the post is on the street
- DESCRIPTION_REP: Description REP of sign
- DESCRIPTION_RTP: Description RTP
- NOM_ARROND: Boroughs of sign
- Longitude:Longitude (WGS84) of post
- Latitude: Latitude (WGS84) of post
- X: X Coordonnates (NAD83 MTM8) of post
- Y: Y Coordonnates (NAD83 MTM8) of post

The description of the fields in the `arrondissements` (boroughs) is as follows

- CODE_ID: Identification number of the borough
- NOM: Borough name
- NOM_OFFICI: Borough official name
- CODEMAMH: MAMH code
- CODE_3C: 3 letter abbreviation
- NUM: Reference Number

- ABREV: 2 letter abbreviation
- TYPE : type (borough or linked city ("ville liée"))
- Aire: Area

Please give the SQL queries to extract the following information:

1. the number of at least two sign categories (1 pt);
2. the list of cities and boroughs sorted by area (0.5 pt);
3. the number of signs per borough and the number of boroughs; (1.5 pts)
   - comment on the results;
4. the number of cities and boroughs and their average area according to their type; (1pt)
5. the number and density (number per unit of area) of signs per borough; (2 pts)
   - comment on the results.

Solution

1. Several fields could be used to define panel categories, such as DESCRIPTION_RPA, DESCRIPTION_CAT, DESCRIPTION_REP and DESCRIPTION_RTP. A possible query is therefore: `SELECT COUNT(DISTINCT DESCRIPTION_RPA), COUNT(DISTINCT DESCRIPTION_CAT)`
   `FROM signaling`
2. `SELECT NAME, Area FROM arrondissements ORDER BY Area`
3. The query for the number of signs per district is: `SELECT NOM_ARROND, count (*) FROM signaling GROUP BY NOM_ARROND`. The query for the number of districts is: `SELECT count(*) FROM districts`.
   - There are thus 34 elements in the districts table, but only 20 results when we count the signs by district in the signaling table (with an additional category of signs not associated with a district). The boroughs table contains cities in the west of the Island of Montreal which are not available in the signaling table.
4. `SELECT type, COUNT(*), AVG(area) FROM arrondissements GROUP BY type`
5. `SELECT S.NOM_ARROND, COUNT(*) AS number, COUNT(*)/A.Aire AS density FROM signalisation S, arrondissements A`
   `WHERE S.NOM_ARROND = A.NOM_OFFICI GROUP BY S.NOM_ARROND`
   - This query only gives 11 results because the names of the districts do not correspond exactly in the two tables, even for the districts having signs, for example "Rosemont - La Petite-Patrie" and "Rosemont-La Petite -Country".