Exam preparation exercises

1 Types of variables

2 Data collection methods

Exercise 1 (periodic 2013) 20 min (/ 3)

Identify three different survey techniques (modes): for each, specify 2 advantages and 2 disadvantages.

Exercise 2 (final 2014) 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.

Solution

- Some advantages of the Origin-Destination surveys carried out in the greater Montreal area: sample size, disaggregated approach (individual trips by users)
- Some disadvantages of the Origin-Destination surveys carried out in the greater Montreal area: period frozen in time, partial or incomplete information (because reported by one person for all the members of the household)

Exercise 3 (quiz 2012)

1. From which list is the sampling frame for the Origin-Destination household survey in the Montreal region made?
2. Give examples of expected biases when carrying out the Origin-Destination household survey in the Montreal region, and the reasons for these biases?
3. What is the expansion factor for?
   (a) That one respondent from an under-represented group in the sample counts as more than another respondent from an over-represented group in the sample
   (b) Weight the sample so that it represents the size of the reference population
   (c) To eliminate the non-response rate
   (d) To visualize more precisely the data collected
   (e) All of the above

Solution

1. The landline list
2. 1) Under-representation of young people who have fewer and fewer landlines; 2) Under-representation of busy families, who do not have time to answer the phone.
3. Response b

Exercise 4 (survey methods) 25 min (/ 4 pts)

The Laurentides Intermunicipal Transport Council (CIT) is the transport company serving a suburban area north of Montreal. You wish to collect information concerning the profile of users of bus line 23, as well as their profile of use of this bus line. The route of the bus line connects the Sainte-Thérèse train station and the municipality of Sainte-Anne-des-Plaines, via the Collège Lionel-Groulx (CEGEP).

1. What is the reference population for this survey?
2. What data collection technique do you suggest? Explain how this collection works.
3. What will be the format of the questionnaire?
4. Identify an appropriate time frame for this survey as well as the time unit. To justify.
5. What is the minimum sample size you need to collect? You want a 95 % confidence level and you accept a 4 % margin of error. Based on smart card transaction data, you know that your benchmark population is 2,000 individuals, and you want to make sure you meet the 75 % proportion of students in the line customer base.

Solution

1. Users who use bus line 23
2. Interception, investigation aboard line 23 buses
3. Paper or iPad
4. A weekday in the fall (during the school and work calendar)
5. The minimum sample size is $n = (1.96^2 \times 0.75(1 - 0.75))/0.04^2 = 450$ (limited pop correction $n' = 450/(1 + 450/2000) = 367$ individuals)

3 Data processing

Exercise 1 (quiz) (
/1 Pt)

Note that each displacement d_i consists of n_i points $(x_{i,j}, y_{i,j})$, where $1 \leq j \leq n_i$, in a Cartesian coordinate system. Write an algorithm that calculates the length of a displacement. Use the pseudo-code format shown in the course, specifying the inputs and outputs.

4 Databases

Exercise 1 (periodic 2012, 2019) 45 min (/ 6 pts)
We are interested in creating an information system for an airline. For this, it is necessary to model the different objects and concepts necessary for its operation. These entities are: airport, flight, plane, employee, type of employee, garage, passenger, ticket.

1. Provide a data model in the form of an Entity / Association diagram involving all the entities listed above. Add attributes (indicating the 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 state the primary and foreign keys (and what the foreign keys refer to), and provide types for the attributes. (2 pts)

Solution

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

   airport id, name, city, number of runways

   flight id, origin, destination, date and time of departure, date and time of arrival

   aircraft id, make, model, date of manufacture

   employee id, name, date of birth

   employee type id, description

   garage id, address, size

   passenger id, name, gender, date of birth

   ticket id, price

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

   - plane-flight: a flight involves 1-1 aircraft, an aircraft is involved in 1-n flights. The functionality is 1-n.
   - flight-airport: a flight involves 2-2 airports (arrival, departure), an airport is the point of departure or arrival of 1-n flights. The functionality is 2-n.
   - flight-ticket: one ticket is used to fly on 1-1 flight, 1-n tickets are sold for one flight. The functionality is 1-n.
   - passenger-ticket: a ticket allows 1-1 passengers to take the flight, a passenger could buy 0-n tickets (in his life). The functionality is 1-n.
   - theft-employee: a theft involves 1-n employees (on the flight), an employee works / participates in 0-m flights (0 if remains on the ground). The functionality is n-m.
   - employee-type of employee: an employee has 1-n type of jobs (historically), a type of job can correspond to 1-m employees. The functionality is n-m.
   - garage-airport: a garage is located near 1-1 airport, an airport can have 0-n garages. The functionality is 1-n.

We could add associations between garage and airplane (the garage typically used by an airplane) and between employee and garage (for employees working in maintenance in a garage) or employee and airport.

2. Each entity becomes a table (Airports, Flights, Planes, Employees, Employee types, Garages, Passengers, Tickets). Two tables must be added to represent the n-m associations, VolEmployés and EmployéTypeEmployés, which are made up of two external keys to the primary keys of the tables concerned by the associations (respectively Flights and Employees, and Employees and Types of employees). The following external keys must be added for the n-m associations:
   - airportDepartureId, airportArrivalId and airplaneId in Flights referring respectively to Airports.id (2) and Airplanes.id;
   - flightId and passengerId in Tickets referring to Flights.id and Passengers.id;
   - airportId in Garages referring to Airports.id.

Exercise 2 (data models) 45 min (/ 6 pts)

We are interested in creating an information system for a shipping company. For this, it is necessary to model the different objects and concepts necessary for its operation. These entities are as follows: port, route, ship, employee, type of employee, home port (single reference port for a ship), container (“ container ”).

1. Provide a data model in the form of an Entity / Association diagram involving all the entities listed above. Add attributes (indicating the 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 state the primary and foreign keys (and what the foreign keys refer to), and provide types for the attributes. (2 pts)

Solution

There are no n-m associations in this E / A diagram. A solution was presented during the course.

Exercise 3 (final 2014) 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 the 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 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 in time. (1 pt)
4. What feature of database management systems protects the individual information of users participating in the project? (0.5 pt)

Exercise 4 (quiz) (/ 2 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 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 journey mobile application in the form of an Entity / Association diagram involving at least the following entities: respondent, movement, GPS point. Add attributes (including the identifier) and associations between entities, with their minimum and maximum cardinalities, and functionalities.

2. Translate the Entity / Association schema into a relational schema. Clearly indicate the primary and external keys.

Exercise 5 (final 2012) 30 min (/ 4)

Present a data model for a transport system. You have two choices: a car and truck rental agency or a centrally managed carpooling service. Your model should consist of at least five entities and form a cohesive whole (which does not lack an important element necessary for the main functionality of the system). Present the relational model for such a system. Clearly indicate:

1. the primary keys;

2. foreign keys;

3. of the relevant attributes;

4. the data types of these attributes;

5. the functionality of each relation;

6. you must have at least one relation of type *many-to-many* (*n-m*).

Exercise 6 (periodical 2010) 45 min (/ 6 pts)

The city of Montreal would like to create an information system to manage public parking spaces in the city. In particular, this system should allow an exhaustive inventory of locations, paid or not, and their use. It is assumed that the paid locations are equipped with a parking money collection device (parking meter) and that the parking cost per hour is fixed over time. When a driver parks in one of these locations, he must pay at the parking meter for a certain period. It should also be noted that some parking lots have specific restrictions, for example some are reserved for disabled people.

1. Provide a data model in the form of an Entity / Association diagram that allows you to record all the information on public parking spaces and their use as described above. Add attributes (indicating the 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 state the primary and foreign keys (and what the foreign keys refer to), and provide types for the attributes. (2 pts)

Exercise 7 (final 2014) 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 `14-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 has an impact on 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 obtain 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-means algorithm to identify homogeneous groups of stations with similar flows according to the time of day. Present the results in a few lines. Represent the centroids of the groups on a figure. (2.5 pts)

Solution

1. The primary key of the " loopdata " table is the composite key (detectorid, starttime). The table follows the first normal form, but not the second because the date, time, and daytype attributes relate only to part of the primary key (starttime).

2. `SELECT daytype, AVG (speed), COUNT (speed) FROM loopdata WHERE detectorid = 1732 GROUP BY daytype`

daytype	AVG (speed)	COUNT (speed)
0	48.8148333333333	120
1	50.1735338345865	133
2	49.0002097902098	143
3	49.8603539823009	113
4	50.9990517241379	116
5	50.5410743801653	121
6	50.1680172413793	116

The appropriate statistical test to determine whether the day of the week has an impact on the average speed at that station is one-way analysis of variance (ANOVA). You can use Excel or Tanagra to do this test. The data must be exported with the following query:
`SELECT daytype, speed FROM loopdata WHERE detectorid = 1732 ORDER BY daytype;.`
The null hypothesis is that the mean of the speeds is identical for all the groups (alternative hypothesis: at least one mean is different). The statistic of the test is $F = 3.13$, which corresponds to a risk of the first kind of 0.0048, which is very low. We can reject the null hypothesis and conclude that the average speed changes depending on the day of the week.

3. `SELECT detectorid, time, AVG (volume) FROM loopdata WHERE daytype BETWEEN 1 AND 5 GROUP BY detectorid, time ORDER BY detectorid, time;`

4.
   (a) Here is the example of the query to create the first view:
   `CREATE VIEW qnight AS SELECT detectorid, AVG (volume) AS volume FROM loopdata WHERE (daytype BETWEEN 1 AND 5) AND (time BETWEEN "18:00:00" and "23:00:00") GROUP BY detectorid;`
   (b) Let qnight, qmorning, qafternoon and qevening be the four views. The request to join the views is as follows: `SELECT qnight.detectorid, qnight.volume as flow_night, qmorning.volume as flow_morning, qafternoon.volume as flow_afternoon qevening.volume as flow_evening FROM qnight, qmorning, qafternoon, qevening WHERE qnight.detectorid = qmorning.detectorid and qmorning.detectori = qafternoon.detectorid and qafternoon.detectorid = qevening.detectorid`
   (c) By choosing three groups, the stations are divided into high, medium and low flows for the four periods considered. The third group of 21 stations has consistently higher average flow rates, while the second group of 18 stations

has consistently lower average flow rates. The first group contains the most stations (30) and shows average flows between the flows of the other two groups for all periods, slightly higher than the overall average for the morning and afternoon and lower for the night. and the party. The centroids are as follows and can be represented on a graph with parallel coordinates (the averages of the flows of each group as a function of the period on the x-axis).

	Group 1	Group 2	Group 3
flow_night	293.059309	146.422476	443.714401
flow_morning	1008.166874	470.038251	1327.814308
flow_afternoon	1096.418644	534.436908	1338.707613
flow_evening	598.190248	327.042045	874.911848

5 Spatial data

Exercise 1 (quiz 2012) Which system is more precise: MTM or UTM? Why?

Solution MTM: each zone corresponds to an angle of 3 °. At the edges of each zone, it is more precise than the UTM zones corresponding to an angle of 6 °.

Exercise 2 (periodical 2014)

1. What is the difference between a geoid, a datum and an ellipsoid? (1 pt)
2. Mercator projection
   (a) What is the main problem with the Mercator projection (the one commonly used to represent the earth on a map). (0.5 pt)
   (b) Where in the world is this the most problematic? (0.5 pt)

Solution

1. A datum is a reference system for expressing positions in the vicinity of the Earth, involving a model of the shape of the earth, usually coordinates in angle units (e.g. degrees), and an origin $(0, 0)$.
   The model of the earth is generally a conventional ellipsoid of revolution (chosen so as to approach the geoid) whose defining parameters are generally its center (chosen near the earth's center of gravity) and three orthonormal axes defined by their orientation.
2. Mercator projection
   (a) The earth is projected onto a cylinder, coinciding with the earth at the equator, and whose distance error increases with distance from the equator.
   (b) Near the poles.

Exercise 3 (final 2017) 45 min (/ 6 pts)

You have the following tables.

- Buroughs table:

Field	Type
id_burough	Integer
name_burough	VARCHAR (255)
Geom	Geometry (MultiPolygon, 32188)

- Road Network table:

Field	Type
id_link_road	Integer
Geom	Geometry (MultiLinestring, 32188)

- Cycle Network table:

Field	Type
id_link_bike	Integer
Geom	Geometry (MultiLinestring, 32188)

Propose a method, for example in the form of an SQL query with spatial functions, in order to determine, by district, the proportion of the road network that contains a cycle lane. The list of spatial functions is presented in the table 1.

Function	Description
ST _Area (g1)	Returns the area of the surface if it is a Polygon or MultiPolygon
ST _Dwithin (g1, g2, distance _of _srid)	Returns true if the geometries are within the specified distance of one another
ST _Intersection (geomA, geomB)	Returns a geometry that represents the shared portion of geomA and geomB
ST _Intersects (geomA, geomB)	Returns TRUE if the Geometries / Geography "spatially intersect in 2D"
ST _Length (g1)	Returns the 2d length of the geometry if it is a linestring or multilinestring
ST _X (g1)	Return the X coordinate of the point
ST _Y (g1)	Return the Y coordinate of the point

Table 1: List of spatial functions

Solution
The steps of the method are as follows:

1. Creation of a table of road links by district:
```
CREATE TABLE public.road_network_burough AS SELECT l.*, r.id_burough,
ST_Intersection (l.geom, r.geom) as geom_intersection FROM public.road_n
l INNER JOIN public.buroughs r ON ST_Intersects (l.geom, r.geom);
```

2. Creation of a table of cycle links by district:
```
CREATE TABLE public.bike_network_burough AS SELECT l.*, r.id_burough,
ST_Intersection (l.geom, r.geom) as geom_intersection FROM public.bike_n
l INNER JOIN public.buroughs r ON ST_Intersects (l.geom, r.geom);
```

3. Extraction of cycle paths which are at a certain distance from the road network (here 10m):
```
CREATE TABLE public.bike_network_rounded_within10m AS SELECT DISTINCT
ON (a.id_link_bike) a.  * FROM public.bike_network_buroughs to,
public.public_road_burough b WHERE ST_DWithin (a.geom_intersection,
b.geom_intersection, 10);
```

4. Calculation of the lengths of the selected cycle network and of the road network by district:

```
CREATE TABLE buroughs_length AS SELECT l.id_burough, sum (ST_Length
(c.geom_intersection)) / sum (ST_Length (r.geom_intersection))
as percentage_bike_network FROM boroughs l LEFT JOIN bike_network_burough
c ON l.id_burough = c.  id_burough LEFT JOIN road_network_burough
r ON l.id_burough = r.  id_burough GROUP BY l.name_burough;
```

6 Statistical analysis

Exercise 1 (quiz 2012) Knowing that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ (where $\bar{X}$ is the empirical mean of n samples of X, μ and σ the mean and the standard deviation of X) tends towards the reduced centered normal distribution, calculate a confidence interval of the mean of 100 samples of speeds of empirical mean 55 km / h and standard deviation 8 km / h. Specify the confidence level of the interval (if Z is a real random variable with reduced centered normal distribution, $P(Z \leq 1.96) = 0.975$ and $P(Z \leq 1.645) = 0.95$).

Solution The 95 % confidence interval is $[53.43, 56.57]$ ($\mu \pm 1.96 * \sigma/\sqrt{n} = 55 \pm 1.96 * 8/10$).

Exercise 2 (quiz 2012) We test the hypothesis H_0 if a speed sample follows the normal distribution: the decision variable of the χ^2 test is calculated and is equal to 14.3574. The number of degrees of freedom is 9 and the threshold values for a distribution of χ^2_9 are 14.68 and 16.92 for risks of the first kind of 10 % and 5 % respectively. Conclude.

Solution The risk is too great (greater than 10 %) to reject the null hypothesis, so we cannot reject the null hypothesis that the velocities follow the normal law. So it seems that the speeds follow the normal law.

Exercise 3 (final 2010)

After widening the lanes on the road studied for another question (data mining section, `speed-flow.csv` file (in `10-vitesse-flow.zip`)), a new reading is carried out of data contained in the file `speed-flow2.csv` (in `10-vitesse-flow.zip`). We would like to know if this development has had an impact on the speeds practiced by drivers on this road. After transforming the data into a number of observations per speed intervals (for example using the histogram function in Excel), indicate whether the arrangement had a significant impact on the distribution of speeds (with a level of confidence of 95 %).

Solution

We must test the hypothesis H_0: the distribution of speeds is identical before and after the arrangement. To do this, the data must be transformed into a number of observations per interval (the Excel histogram tool constitutes 15), by grouping the intervals with less than 5 observations. Taking the data collected before as a reference, we calculate the decision variable of the χ^2 test: the value obtained is 63.18, greater than the value of 14.07 corresponding to a confidence level of 95 % for a random variable following a law of χ^2_7

at 8-1 = 7 degrees of freedom. We can therefore conclude that the development had a significant impact on the speed distribution on this road.

Exercise 4 (periodic 2013) 50 min (/7 pts)

The number of accidents on a road is counted in the following table for 15 days (period 1):

Day	Number of accidents
1	1
2	1
3	1
4	0
5	2
6	0
7	0
8	0
9	2
10	0
11	1
12	1
13	1
14	1
15	1

1. Write the algorithm for calculating the median of a set of n real numbers x_i. (1 pt)
2. Calculate the average and the median of the number of accidents per day. (1 pt)
3. Calculate a 95 % confidence interval for the average number of accidents per day. (1 pt)
4. Calculate the number of days of observation necessary to obtain the average number of accidents per day with an accuracy (tolerance) of 0.15 accidents per day for a confidence level of 90 and 95 %. (1 pt)
5. Draw the histogram of the distribution of the number of accidents per day (and not the time series of the number of accidents as a function of the day). (1 pt)
6. To improve road safety, a policeman is placed visibly on the side of the road for 15 days. During this period 2, the average number of accidents per day is 0.45 and the empirical standard deviation has not changed (it is assumed that the variances are the same for periods 1 and 2 and that the number of accidents per day follows a normal law). Determine if the number of accidents has decreased with a risk of error of the first kind of 5 %. (2 pts)

Solution

1. Here is an algorithm (assuming an existing *sort* sorting function on real numbers, current view):

input: n real numbers x_i
output: the median of the n real numbers x_i

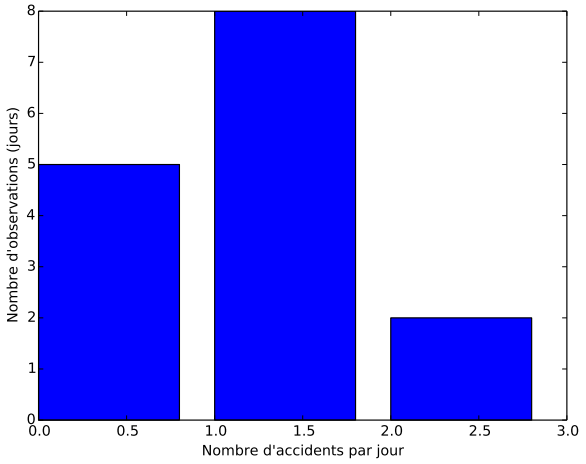
start
$sorted\_list = sort(x_i)$
if n even
return the element in position $n/2$ of $sorted\_list$
otherwise
return the element in position $(n-1)/2$ of $sorted\_list$
end

2. The mean and median of the number of accidents per day are 0.80 and 1 accidents per day respectively.

3. The expression $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ follows Student's law with 14 degrees of freedom, and the probability that such a variable is respectively in the interval $[-2.145, +2.145]$ and $[-1.761, +1.761]$ is 95 % and 90 %. The corrected standard deviation s is 0.68 and the confidence interval of the average number of accidents per day is therefore respectively $0.8 \pm 2.14\frac{0.68}{\sqrt{15}} = [0.42, 1.18]$ and $[0.49, 1.11]$ for confidence levels of 95 and 90 %.

4. We assume that the empirical standard deviation is close to the true standard deviation. The number of observations required is respectively $n = 1.64^2\frac{0.68^2}{0.15^2} = 55$ and $n = 1.96^2\frac{0.68^2}{0.15^2} = 79$ for confidence levels of 90 and 95 %.

5. The histogram of the distribution law of the number of accidents per day below is obtained by the Python code at the end of the exercise:



6. We test the hypothesis H_0: the average number of accidents has not changed against H_1: the number of accidents has decreased. The test statistic is $\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = (0.8-0.45)/(0.68\sqrt{2/15}) = 1.41$ ($n_1 = n_2 = 15$, $s_1 = s_2 = 0.68$). The statistic follows Student's law at $n = 15 + 15 - 2$ degrees of freedom. The cut-off value of

the distribution for a risk of the first kind of 0.05 is 1.701 (ie the probability that a variable according to Student's law with 28 degrees of freedom is greater than or equal to 1.701 is 0.05). We cannot therefore reject H_0, the number of accidents does not seem to have been affected. We can find in the table that the value p (or risk of the first kind) for 1.41 is between 0.10 and 0.05, which could be accepted with a confidence level of 90 %.

Exercise 5 (final 2013) 55 min (/ 8 pts)

This exercise is based on a set of 3000 accidents involving a pedestrian and a vehicle between 2003 and 2006 in the city of Montreal (the file `13-accidents-pietons-montreal.txt` is available on moodle). The data is in the form of a text file (with the fields separated by a tab), and each accident is described by the attributes described in the table 2.

Attribute	Description
EVENT	accident number
RDCLASS	road classification (1: motorway; 2: numbered road; 3: collector; 4: artery; 5: local)
SPD_KM	speed limit according to road classification
MED_INC	median income in the accident area
pop_dens_200	population density within 200 m
veh_type	type of vehicle (" car ": car; motorcycle; " VTB ": van, truck (" truck ") or bus; " EMS ": emergency vehicle))
BAD _WEAT	bad weather indicator variable
SEVERITY	severity of the accident (3: fatal; 2: serious injury; 1: slight injury; 0: no injury)
DARK	variable for the night
Park_10	presence of a park 10 m from the accident
hosp_50	presence of a hospital within a radius of 50 m
veh_mvt	movement of the vehicle involved (" straight "; " backup "; " leftturn "; " righturn ")
Comm_Per	percentage of business activity
Res_Per	percentage of residential activity
Inter_Acc	occurrence of the accident at a crossroads

Table 2: Accident attributes

1. Discuss statistical models that can be used to study the association of these attributes with accident severity. (1 pt)
2. Describe the processing required to use nominal data in a regression analysis (eg linear regression). (1 pt)
3. By creating a new binary variable representing fatal and serious injury accidents (the variable is 1 if the accident is fatal or with serious injury, 0 otherwise), choose a statistical model to study the factors contributing to the probability of a fatal or serious accident: estimate the model (with Tanagra), clearly present the significant attributes and discuss the results. (4 pts)

4. Draw a histogram of the distributions according to the road class of the number of fatal and serious accidents on the one hand, and the number of accidents with minor injuries and without injuries on the other hand: apply a statistical test to determine whether the two distributions are different. (2 pts)

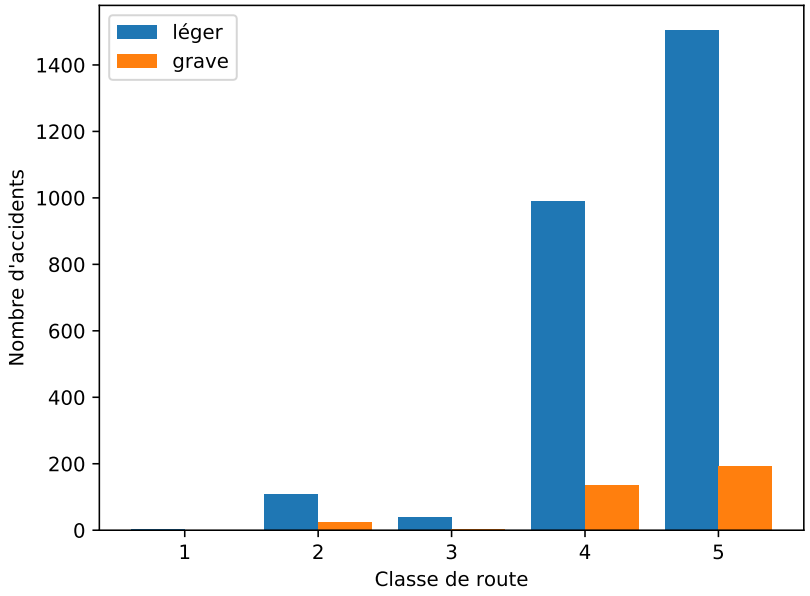
Solution

1. The severity of an accident is represented by an ordered nominal variable. An ordered logit model is the most suitable for studying the association between the attributes of accidents and their severity (dependent variable). A multinomial model could also be used, but would not use the severity level order information.

2. Nominal data taking K values where $K \geq 3$ must be replaced by $K - 1$ binary variables representing each of the values taken (for example, the class variable of the route will be represented by four variables for highways, numbered roads, collectors and arteries, local streets being represented by the zero (false) value of these four binary variables).

3. We create the binary variable " severity0 " to represent fatal accidents and accidents with serious injuries: the variable is equal to 1 if the accident is fatal or with serious injuries, 0 otherwise (for Tanagra, it may be advantageous to use text for the values of this variable so that it is directly recognized as categorical (binary)). A sample Tanagra file is provided. We see in the model that for example the variable of the speed limit is significant (confidence level of 95 %) and negatively correlated, which corresponds to knowledge of road safety (the higher the speed, the more an accident is serious).
   Answer to be completed.

4. The statistical test is the χ^2 test that compares two samples of data. You have to choose a reference sample, normalize it and multiply by the number of observations of the other sample to obtain the expected numbers. The histogram of the numbers of accidents according to the type of road is as follows:



The table for the χ^2 test considering the number of minor accidents as a reference is as follows (after grouping together the road categories for which there are less than five observations):

Road class	Number of light accidents	Number of serious accidents
1 and 2	179.39	112
3	22.42	39
4	1016.54	991
5	1427.64	1504

The null hypothesis of the test is that the distributions are identical. Under the null hypothesis, the test statistic follows the law of χ^2 at $d = n - 1 - p = 4 - 1 - 0 = 3$ degrees of freedom. The statistic of the test is $X^2 = 42.29$, which corresponds to a risk of the first kind less than 10^{-6}. We reject the null hypothesis, the distributions of light and serious accidents according to road classes are different. It can be concluded that the severity of accidents is not the same on the different classes of roads.

7 Regression and econometric modeling

Exercise 1 (final 2014) 30 min(/4.5 pts)

This exercise is based on the data set of car characteristics contained in the file `autos.txt`. 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 the width as a function of the length and calculate the correlation coefficient: comment. (1 pt)

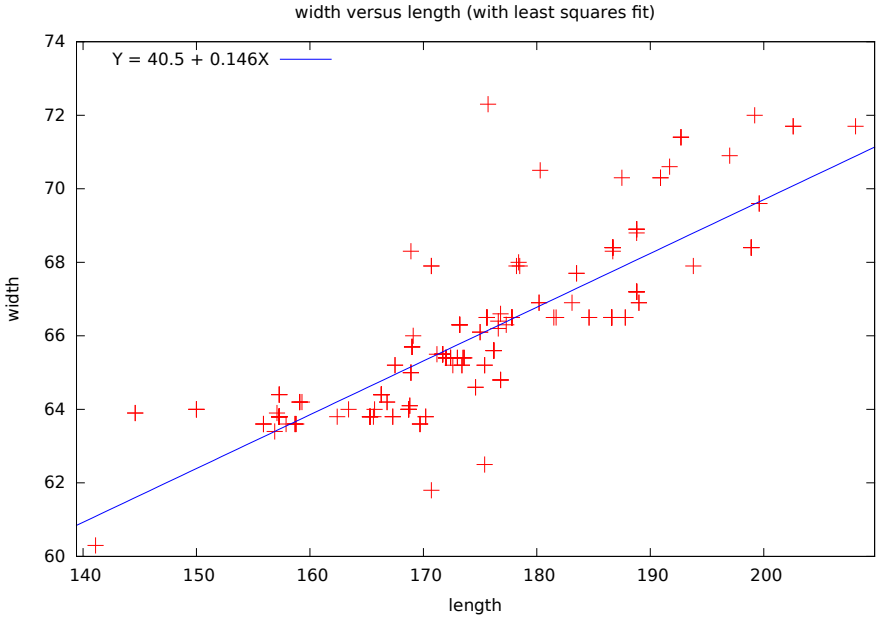
2. Using one of the software at your disposal, estimate the linear regression line of the width as a function of the length:

(a) Discuss the significance of the model and calculate (without using Excel) 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 $\hat{}$ 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)

Solution

1. The scatter plot for width versus length is as follows:



The linear correlation coefficient between the two variables is 0.841, which is very high. The wider a vehicle, the longer it is (in this dataset).

2. (a) The model parameters are $\hat{a} = 0.146253$ and $\hat{b} = 40.452511$. The model is significant (very low risk of the first kind, of the order of 10^{-56}). The confidence interval of a is $[\hat{a} - t_{\alpha/2} s_{\hat{a}}, \hat{a} + t_{\alpha/2} s_{\hat{a}}]$ where $t_{\alpha/2}$ is such that $P(|t| < t_{\alpha/2}) = 1-\alpha$ for a random variable t following the Student's law at 203 degrees of freedom (Student's law tends towards the normal distribution when the number of degrees of freedom becomes large). We find $[0.1353, 0.1571]$ and $[0.1332, 0.1592]$ respectively for the confidence levels of 90 % and 95 %.

(b) The linear regression assumptions concerning the residuals seem to be respected since the residuals are equally distributed on either side of the x-axis (zero mean and constant variance). There are some outliers with larger errors.

8 Data visualization

9 Data mining and machine learning

Exercise 1 (final 2010) 35 min (/ 5 pts)

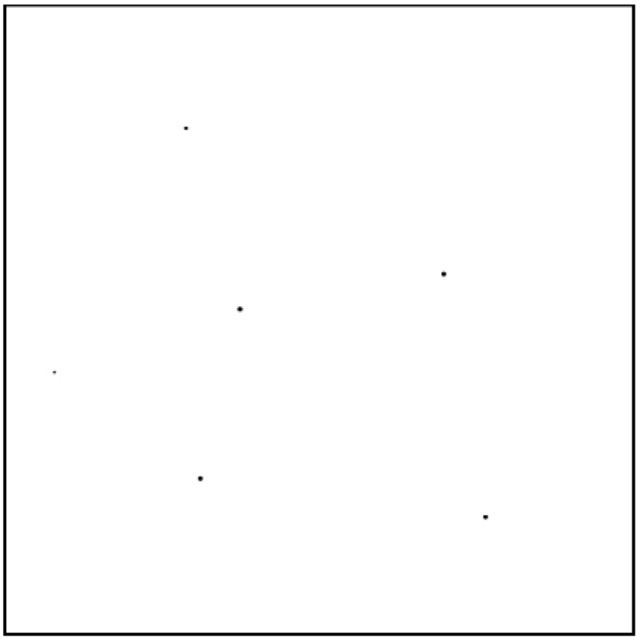
The file `speed-debit.csv` (in `10-vitesse-debit.zip`) contains observations of speeds (in km / h) and flows (in number of vehicles per hour) by interval of 15 min for a direction of a two-lane rural road.

1. Describe two methods, one intrusive and one non-intrusive, of collecting speed data on a road (name one advantage and one disadvantage for each). (0.5 pt)
2. After visual exploration of the data, propose a distribution of the data into " homogeneous " groups: justify your choices, characterize these groups by their summary descriptive statistics and propose a short qualitative description of the groups. (2 pts)
3. Using the first ten observations in the file as an example, illustrate in a few steps (at least 3 steps, including initialization and final step) how a data segmentation algorithm works. Discuss the treatment (s) required prior to segmentation. (2 pts)
4. Indicate how it would be possible to represent a third variable (for example the proportion of heavy goods vehicles per interval of 15 min on this road) in a cloud of points in the space of speeds and flows. (0.5 pt)

10 Spatial analysis

Exercise 1 (periodic 2013) 45 min (/ 6 pts)

1. Give an example of a centrality measure in spatial analysis and explain what this can be used for in a transport context. (1 pt)
2. Give an example of a dispersion measurement in spatial analysis and explain what this can be used for in a transport context. (1 pt)
3. What are the overall Moran I Index and the Global Geary C Index used for? In what context could they be used in transport? (1 pt)
4. Is it possible to obtain a projection of the earth on a plane which preserves distances, shapes, angles and areas? Why? (1 pt)
5. Draw what the Voronoi diagram would look like (according to Euclidean distance) around the points below (don't measure exactly, only sketch). (2 pts)



Exercise 2 (periodical 2014) 30 min (/ 3 pts)

1. Distribution types:
   (a) Draw what a point process generated dot pattern would look like with a totally random spatial structure in the frame below. (0.5 pt)

   (b) What would be the value of the Moran index I and the Geary index C for this distribution? (0.5 pt)
2. Give two relevant examples of the use of Thiessen polygons (or Voronoi diagrams) in the field of transport. (1 pt)

Solution

1. Distribution types:
   (a) Draw points according to the uniform law independently for the two coordinates.
   (b) $I = 0$ and $C = 1$.
2. For example, polygons are used to convert point data into zonal data, to estimate the target population for a bus stop, to delimit the borders of an area for a household.