Periodic check

J.-S. Bourdeau and N. Saunier

October 4, 2016

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;
- download the data necessary for certain exercises available in the data archive on moodle (Section " Periodic control ", first element " Data ") (the text files are provided in a version with a period and a comma for decimals, if necessary).

Statistical tables are included at the end of the statement. If you wish, a file (Excel or Word document) can be placed on moodle for questions involving carrying out calculations from the data provided for certain exercises.

Exercise 1 (survey methods) 30 min (/ 5)

1. Identify three different survey methods (survey techniques): for each, specify two advantages and two disadvantages. (3 pts)
2. Define the respondent's burden and propose three tools, methods, mechanisms or strategies that make it possible to reduce it during a survey on mobility. (1 pt)
3. From the distribution of the population by age group taken from the 2013 origin-destination survey of the greater Montreal area below, calculate the mean and the standard deviation, then the size of the 'sample necessary to have a precision (tolerance) of 1 year on the mean age of the population with a confidence level of 99 %. (1 pt)

0-19 years	20-44 years	45-64 years	65 years and over
975000	1459000	1215000	639000

Solution

1. Here are some examples of methods with advantages and disadvantages:
   - Auto-fill: postal (send and / or return)

     advantages cheaper, large geographic coverage, flexibility for the respondent

disadvantages increased non-response rate, no direct quality control

- Telephone:

  advantages large geographical coverage, centralized supervision of interviewers, CATI, economical, allows validation and clarification of questions / answers, allows multi-language interviews BUT limited duration

  disadvantages often only one respondent, more and more call filtering, bias (no landline), sampling frame (# phone)

- Interception: held at a particular site (away from home), when students carry out an activity, survey on board a public transport vehicle, OD cordon survey (road), trip generator
- " Face-to-face ": the interviewer is present to collect the answers

  advantages better response rates, flexibility of the types of questions, allows information to be provided to respondents, faster (data available " instantly ")

  disadvantages expensive, small sample

- In groups: especially for qualitative surveys, between 7-10 people questioned simultaneously (exchange of experiences, attitudes)

  advantages similar to face-to-face, allows a better understanding of the decision-making process

  disadvantages expensive, small sample, responses influenced by interactions

2. Respondent burden is the effort required of the respondent during the survey, which increases the probability that he / she will drop out of the survey before completing it. Ways of reducing it are as follows: classify the answers a priori, reduce the length of the questionnaire, avoid ambiguous questions, use pleasant interfaces and allow you to resume your answers later (on the Internet), etc.

3. Using the average age of each interval and 65 years for the last interval (other choices were possible for the latter), the average age of the population is 38.32 years and the standard deviation s of 19.68 years old. We deduce the sample size to have a tolerance e of 1 year on the mean age of the population with a confidence level of 99 % $N = z^2_{0.995}\frac{s^2}{e^2} = 2.58^2\frac{19.68^2}{1.00^2} = 2578$.

Exercise 2 (data models) 45 min (/5.5)

We are interested in creating an information system to manage car parking in the city of Montreal.

1. Propose a data model in the form of an Entity / Association diagram involving at least the following five entities: car, parking space, parking lot (exterior, storey, underground), parking rate (variable over time) , Street. Add attributes, including identifiers, and associations between entities with their minimum and maximum cardinalities. (2.5 pts)

2. Translate the Entity / Association schema into a relational schema. Clearly indicate primary and external keys, and suggest types for attributes. (2.5 points)

3. Write an SQL query to display the parking spaces in the sement Ville-Marie district. (0.5 pt)

Solution

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

   car registration number, make, model

   parking space id, length, width

   parking lot id, address, type

   price id, day, start time, end time, start date, end date, price per 15 min

   street id, name, buroughs, city

   The associations are as follows:

   - car-place: n to m (historically)
   - place-lot: n to 1 (can be 0 if the parking space is on the street)
   - place-price: n to m (price varies over time for the same place)
   - place-rue: n to 1 (can be 0 if the place is in a parking lot)
   - park-street: n to 1

2. Each entity becomes a table (Cars, Squares, Parks, Tariffs and Streets). It is necessary to add two tables to represent the associations n-m, CarPlace and Place-Tarif, which are made up of two, external keys to the primary keys of the tables concerned by the association. External keys must be added: parcId in Places referring to Parcs.id, rueId in Places referring to Streets.id, rueId in Parks referring to Streets.id. The types of the id would be numeric (integer), price, length and width also numeric (real number), day of the text (1 among the 7 possible choices), the times and dates their respective type, name, district and city of the channels of characters.

3. The query would be:

   - for spaces in a parking lot: SELECT Pl. * FROM Places Pl, Lots Lo, Streets S WHERE Pl.parcId = Lo.id AND Lo.streetId = S.id AND S.buroughs = Ville-Marie
   - for street spaces: SELECT P. * FROM Places P, Streets S WHERE P.streetId == S.id AND S.buroughs = Ville-Marie

Exercise 3 (statistics) 75 min(/9.5)

A speed data file was extracted from a 20-minute video sequence recorded at the intersection of Maisonneuve Boulevard and Ste-Catherine Street on a fall morning in 2015. Users were manually classified and their average speeds were obtained. calculated. The +GP030004-classes-speeds.csv + file contains the following information:

id unique identifier of each user

road_user_type category of each user (1: motorized vehicle; 2: pedestrian; 4: cyclist)

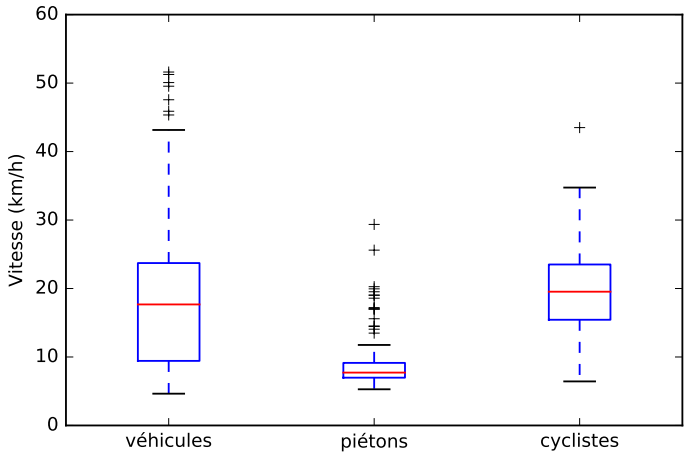
speed average speed of each user (km / h)

1. Calculate the necessary descriptive statistics and plot in the examination book the boxplot for the speeds of each category of road user (for the individual points outside the ends of the box , simply report the number of observations, but do not plot them). Are the distributions symmetrical? Justify the answer. (4 pts)
2. Does the distribution of motorized vehicle speeds follow the normal law? Justify the answer without doing a statistical test. (0.5 pt)
3. Perform a test of the adequacy of the speed distribution of the cyclists to the normal law (using speed intervals of width of 5 km / h). (2.5 pts)
4. Calculate the confidence interval of the average pedestrian speed with a confidence level of 90 % and 95 %. (1 pt)
5. Pedestrian speed data was collected at night: 107 pedestrians were observed, with an average speed of 8.26 km / h and a standard deviation of 4.6 km / h. Use a statistical test to determine whether pedestrians walk slower at night. (1.5 pts)

Solution

1. The vehicle and pedestrian speed distributions are not symmetrical in view of the position of the first and third quartiles on either side of the median. The box whisker for cyclists' speeds seems to show a symmetrical distribution. The speed statistics and boxplots for the different user categories are presented below.

User types		Speed (km / h)
Motorized vehicles	number of observations	598
	average	18.04
	standard deviation	9.25
	minimum value	4.64
	1st quartile	9.43
	median	17.67
	3rd quartile	23.72
	maximum value	51.61
Pedestrians	number of observations	128
	average	9.16
	standard deviation	4.06
	minimum value	5.28
	1st quartile	6.98
	median	7.72
	3rd quartile	9.14
	maximum value	29.36
Cyclists	number of sightings	75
	average	19.39
	standard deviation	6.88
	minimum value	6.43
	1st quartile	15.44
	median	19.53
	3rd quartile	23.51
	maximum value	43.50



2. The vehicle speed distribution probably does not follow the normal law because it is not symmetrical.

3. The null hypothesis H_0 is that the speeds of the cyclists follow the normal distribution. By considering a normal law of average 19.38 km / h and standard deviation 6.88 km / h, we obtain the following table of the numbers of observations in each interval of length 5 km / h, after grouping the intervals counting less than 5 observations (above 30 km / h):

Intervals speed (km / h)	Number of observations actual	theoretical
$[5, 10]$	7.0	6.36
$[10, 15]$	11.0	13.17
$[15, 20]$	23.0	20.64
$[20, 25]$	19.0	19.39
$[25, 30]$	15.0	15.44

The statistic of the χ^2 test is 0.713, while 5.99 is the value such as the probability that a random variable following the law of χ^2 with 2 degrees of freedom (5 intervals of values, minus 1, minus 2 parameters for the normal distribution) is less than or equal to this value is 0.95 (if X follows the law of χ^2 with 2 degrees of freedom, $P(X \leq 5.99) = 0.95$, and $P(X \leq 0.712) = 0.30$). We conclude that we cannot reject the null hypothesis (the risk of the first kind is 70 %) and the speeds of the cyclists seem to follow the normal law.

4. The confidence intervals of the average pedestrian speed with a confidence level of 90 % and 95 % are respectively $9.16 \pm 1.64 \frac{4.05}{\sqrt{128}}$ or $[8.57, 9.75]$ and $9.16 \pm 1.96 \frac{4.05}{\sqrt{128}}$, or $[8.46, 9.86]$ (in km / h). Note that Student's law is used since the standard deviations are estimated from an empirical sample.

5. The null hypothesis H_0 is that the speed of pedestrians is the same at night and during the day and the alternative hypothesis is that it is lower at night. The test statistic is $Z_0 = \frac{9.14 - 8.26}{\sqrt{4.05^2/128 + 4.60^2/107}} = 1.577$. We can use the normal approximation because of the sample size and under the null hypothesis, the test statistic follows

the reduced centered normal distribution. The test is one-sided (the alternative hypothesis is that the speed is lower at night). The risk of the first kind of more than 5 % (the probability that a variable according to the normal distribution is less than or equal to Z_0 is 0.9426) and we cannot therefore reject the null hypothesis that the speed is the same at night as during the day.