Midterm Exam

P.-L. Bourbonnais, J.-S. Bourdeau and N. Saunier

October 20, 2015

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.

Statistical tables are included at the end of the statement.

Exercise 1 (survey methods) 35 min (/5.5)

The Borough of Saint-Laurent wants the Société de transport de Montréal (STM) to improve their service offer in their industrial district in order to better accommodate the employees of different companies. As the sample was not sufficient in the Origin-Destination survey to carry out representative analyzes, the borough and the STM decided to start a specific data acquisition project.

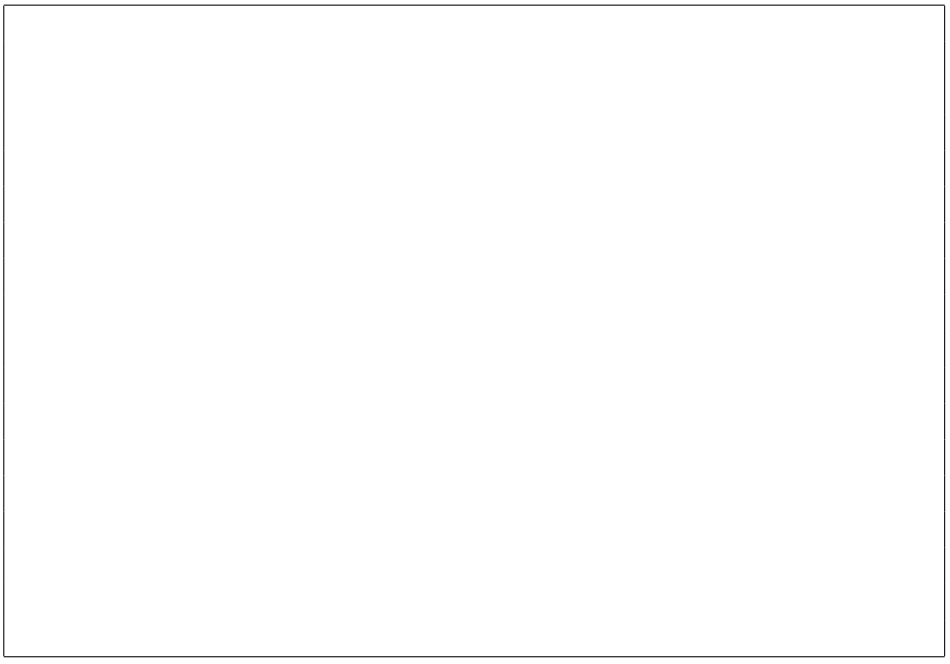
1. Describe the difference between an observation system and a survey. (0.5 pt)
2. What would be the reference population? (0.25 pt)
3. If organizations favor a survey as a data acquisition method, what would be the best survey technique in order to reach the reference population? Explain how it works. (1 pt)
4. Suggest a time frame for the survey. Explain your choice. (0.5 pt)
5. Discuss the sampling rate that would be targeted with regard to the survey technique chosen. (0.5 pt)
6. Name 8 important variables to obtain from the survey. (1 pt)
7. What would be the preferred sampling method? (0.25 pt)
8. Name an example of sampling bias your sample might have as a result of data collection. Explain. (0.5 pt)
9. Assuming that the STM wants to ensure that it accurately estimates (error less than 5 %) the modal share of public transit to Anjou, what should be the minimum sample size to be obtained knowing that the reference population is 2500 people? The OD survey determined that the modal share of public transit to Saint-Laurent was 25 %. (1 pt)

Solution

1. An observation system seeks to observe users (generally without their knowing it to avoid influencing behavior) for the general purpose of collecting data, often for the management of transport systems, while 'a survey consists of directly asking questions to users about their behavior for a specific objective.
2. Employees of borough businesses.
3. A section survey (at a specific period) among employees of the borough's businesses, by telephone or self-completed questionnaire (Internet version available) with recruitment through businesses (and their lists of employees).
4. A period of the year: a typical reference period is autumn (high rate of economic activity).
5. The sampling rate is not that important, what is important is the raw number of responses to estimate the characteristics of the reference population.
6. Age, gender, has a driver's license, place of residence, company, type of job, place of work; for the last day of travel between home and work: mode of transport for travel between home and work, travel time
7. Simple or stratified random (by sector of activity or zones of the borough).
8. An example would be a bias for certain companies or types of companies depending on the ease of reaching their employees and their variable response rate.
9. The sample size is $N = \frac{1.96^2 \times 0.25(1-0.25)}{0.05^2} = 288.1 \approx 289$ people without correction for the population size, or $\frac{289}{1+289/2500} = 259.0$ people after correction.

Exercise 2 (GIS and spatial analysis) 30 min (/ 4)

1. Distribution types:
   (a) Draw what an agglomerated distribution of dots would look like in the box below. (0.5 pt)



(b) What would be the range of values 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 business logistics. (1 pt)

3. What is a geoid? (1 pt)

4. 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. Distribution types:

   (a) The points are grouped in clusters.

   (b) $I > 0$ and $C < 1$.

2. For example, polygons can be used to estimate the potential customers of a store or help in the selection of stores that should be associated with each warehouse.

3. The geoid is the surface coinciding exactly with the surface of the oceans if they were at rest and extended over (under) the continents (calculated using the exact value of gravity at each point on the globe).

4. Mecator projection

   (a) The earth is projected onto a cylinder, coinciding with the earth at the equator, so the distance error increases with distance from the equator.

   (b) Near the poles.

Exercise 3 (data models) 25 min (/ 3)

We are interested in creating an information system for a car rental company. Propose a data model in the form of an Entity / Association diagram involving at least the following entities: car, customer, rental agency, employee. Add attributes, including identifiers, and associations between entities with their minimum and maximum cardinalities. (3 pts)

Solution

The entities and their attributes are as follows:

car registration number (id), make, model, year of manufacture

client driver's license number (id), date of birth, gender, loyalty number

rental id (invoice), start date, end date, price per day, insurance costs

agency id, company, address, car parking capacity

employee id, name, date of birth

The associations are as follows

- rental-client: n to 1
- car-rental: 1 to n
- agency-car: n to 1
- agency-employee: 1 to n
- employee-rental: n to 1

In fact, the rental entity can also be seen as an association n to m between car and customer (or even n-m-l between car, customer and employee).

Exercise 4 (statistics) 60 min (/7.5)

A traffic report was taken on rue des Bois Francs in Victoriaville on August 14, 2012 between 6:00 a.m. and 7:00 p.m.: the result is recorded in a +speed-count.xls + file and provides, among other things, the speeds of each vehicle (column " Speed ") and the slots between each successive vehicle (" Gap Time " column). Please

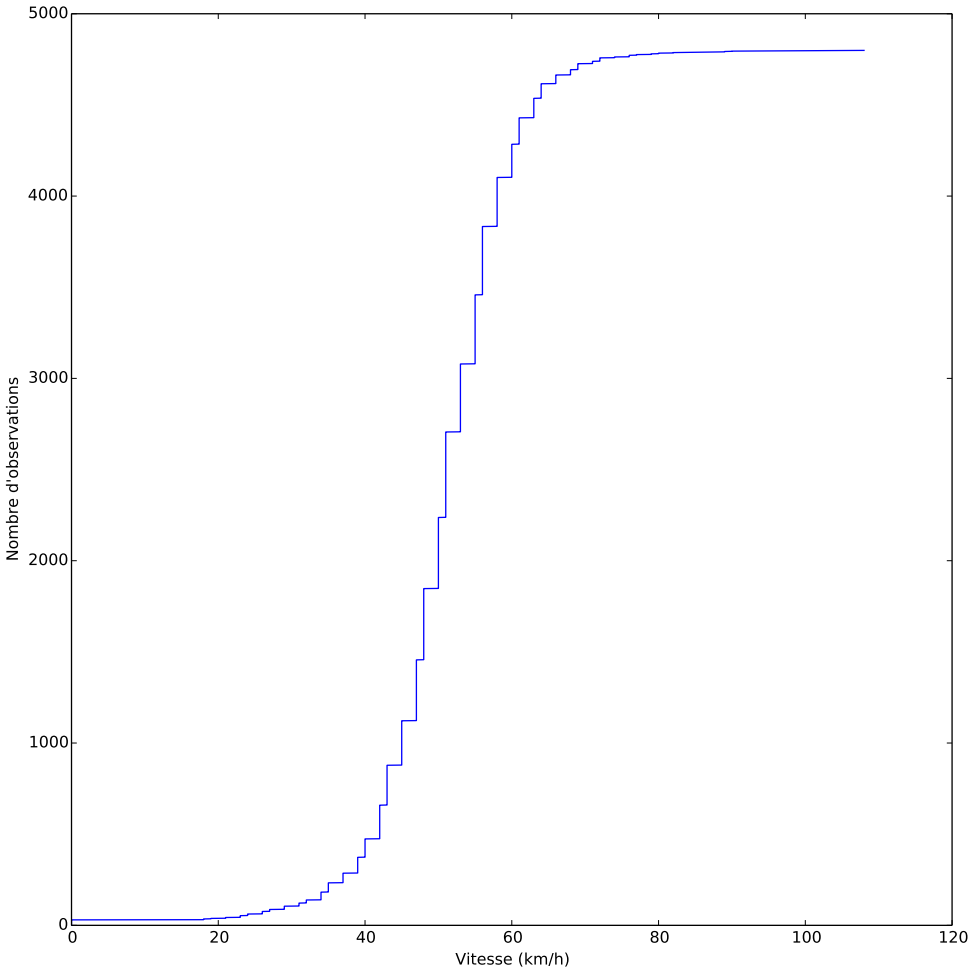
1. Present the cumulative distribution of speeds (in an Excel file to upload to moodle): indicate the quartiles of the distribution (1 pt)
2. Perform a test of the adequacy of the speed distribution to the normal law (using 5 km / h width speed intervals). (2.5 pts)
3. By considering a data table accessible by an index with fields or attributes corresponding to the columns of the file (for example $data[i].speed$ for the (i + 1) th speed measurement or $data[i].t$ the instant of the (i + 1) th measurement), propose an algorithm to calculate the average of the speed between two hours t_1 and t_2. Indicate its complexity according to the number of data (lines). (1.5 pts)

4. Perform a statistical test to determine if the average speed of the morning peak (7:00 a.m. to 9:00 a.m.) is significantly different from the evening peak (5:00 p.m. to 7:00 p.m.) (2.5 pts)

Please include in the same Excel file (or equivalent) the calculations performed for the statistical tests.

Solution

1. The cumulative (non-normalized) velocity distribution is shown in the following figure:



The first quartile, the median and the second quartile are respectively 46.7 km / h, 51.0 km / h and 56.3 km / h.

2. Using intervals of length 5 km / h, from 0 to 110 km / h, then grouped so that there are at least 5 observations per interval (for observed data *and* theoretical) , we obtain 12 speed categories with numbers greater than 5 and the χ^2 test statistic of 691.03 (compared to a normal distribution of average 50.80 km / h and standard deviation 9.38 km / h). The probability that a variable according to the law of χ^2 with 9 degrees of freedom is greater than or equal to this value is 0.0 (risk of the first kind): we can therefore reject the null hypothesis that the distribution of speeds follows a normal distribution (despite the apparent good visual fit).

3. Here is an algorithm:

entry: n vehicle passage records $data[i]$, times t_1 and t_2
output: the average of the speeds between t_1 and t_2

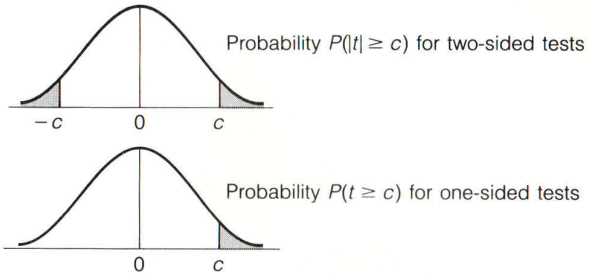
start
$mean = 0.$
$i = 0$
$m = 0$
while $i < n$
if $data[i].t \geq t_1$ AND $data[i].t \leq t_2$
$average = average + data[i].speed$
$m = m + 1$
$i = i + 1$
if $m > 0$
return $average/m$
otherwise
return $None$ (no result)
end

The complexity of the algorithm is linear according to the number of data n, that is to say $O(n)$.

4. We test the hypothesis H_0: the average speed is the same for the two peak periods ($\bar{x}_1$ and $\bar{x}_2$ average morning and evening speeds) against H_1: the average speed is different. The test statistic is $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = (53.23 - 51.94)/(\sqrt{9.20^2/642 + 9.48^2/826}) =$ 2.62. Since there are enough observations, the statistic follows the reduced centered normal. The probability that a normal variable is less than -2.62 or greater than 2.62 is 0.009, which is a very low risk of the first kind: we therefore conclude that the average speed is significantly different between the two peaks in the morning and in the evening (lower in the evening).

Table 3

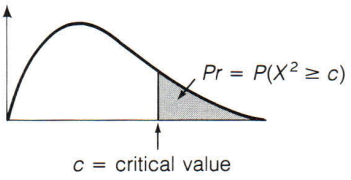
Critical values for Student's t distribution



	PROBABILITY									
	.50	.20	.10	.05	.02	.01	.005	.002	.001	TWO-SIDED TESTS
ν	.25	.10	.05	.025	.01	.005	.0025	.001	.0005	ONE-SIDED TESTS
1	1.000	3.078	6.314	12.706	31.821	63.637	127.32	318.31	636.62	
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598	
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924	
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610	
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	
7	.711	1.415	1.895	2.365	2.998	3.499	4.020	4.785	5.408	
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041	
9	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781	
10	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.537	
11	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437	
12	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318	
13	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221	
14	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140	
15	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	
16	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015	
17	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965	
18	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922	
19	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883	
20	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850	
21	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.257	3.189	
22	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792	
23	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767	
24	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745	
25	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725	
26	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707	
27	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690	
28	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674	
29	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659	
30	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646	
40	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551	
60	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460	

Table 2

Critical values for the chi-square distribution



	Pr							
ν	.500	.250	.100	.050	.025	.010	.005	.001
1	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.83
2	1.386	2.773	4.605	5.991	7.378	9.210	10.60	13.82
3	2.366	4.108	6.251	7.815	9.348	11.34	12.84	16.27
4	3.357	5.385	7.779	9.488	11.14	13.28	14.86	18.47
5	4.351	6.626	9.236	11.07	12.83	15.09	16.75	20.52
6	5.348	7.841	10.64	12.59	14.45	16.81	18.55	22.46
7	6.346	9.037	12.02	14.07	16.01	18.48	20.28	24.32
8	7.344	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	8.343	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.342	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.79	29.82	34.53
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	20.34	24.93	29.62	33.67	35.48	38.93	41.40	46.80
22	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	69.33	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	79.33	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	89.33	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	99.33	109.1	118.5	124.3	129.6	135.8	140.2	149.4

Source: Abridged from Table 8 of *Biometrika Tables for Statisticians*, Vol. 1, edited by E. S. Pearson and H. O. Hartley (London: Cambridge University Press, 1962).

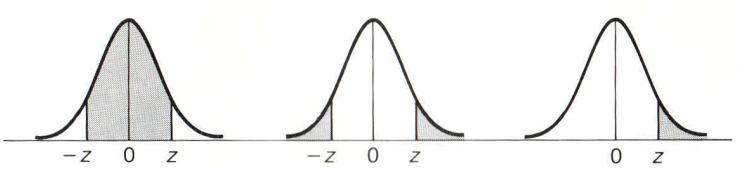


Table 1B

Critical values for the standard normal distribution

CONFIDENCE INTERVALS $P(\|Z\| \le z)$	TWO-SIDED TESTS $P(\|Z\| \ge z)$	ONE-SIDED TESTS $P(Z \ge z)$	CRITICAL VALUE z
.10	.90	.45	.126
.20	.80	.40	.253
.30	.70	.35	.385
.40	.60	.30	.524
.50	.50	.25	.674
.60	.40	.20	.842
.70	.30	.15	1.036
.80	.20	.10	1.282
.90	.10	.05	1.645
.95	.05	.025	1.960
.98	.02	.01	2.326
.99	.01	.005	2.576
.995	.005	.0025	2.807
.999	.001	.0005	3.290
.9995	.0005	.00025	3.480
.9999	.0001	.00005	3.890
.99999	.00001	.000005	4.420
.999999	.0000001	.0000005	4.900

Source: D. B. Owen and D. T. Monk, *Tables of the Normal Probability Integral*, Sandia Corporation Technical Memo 64-57-51 (March 1957).