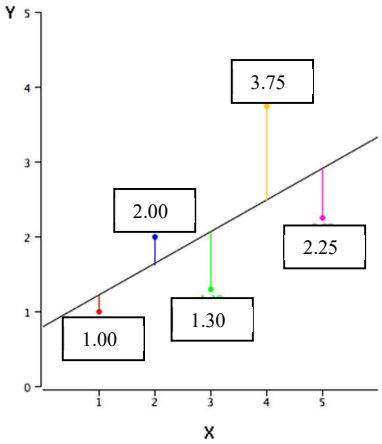
STATISTICAL MODELING AND DESIGN OF EXPERIMENTS (SMDE) -MIRI- (FIB- UPC)

COURSE 2019-2020 Term 1 –FINAL TEST

(Date: 20/01/2019 at 10:00-12:00)

Name:	
DNI or PASSPORT:	
Professors:	Pau Fonseca i Casas
Marks on Racó:	22 of January
Revision:	23 of January on C5218 from 15 to 18

1 Problem (2.5 points): the line



On a dataset you want to analyze, you obtain a chart like this.

We want to approximate y using x values and to do so we expect to find the line that minimized the distances between all the points and the candidate solution.

Can you define the numerical expression of the line that is drawn on this picture?

2 Problem (2.5 points): the new algorithm

A new algorithm based on deep learning methods seems that is working better depending on the meaning of the data contained on the dataset. We are mainly focused on the analysis of Environmental, Social and Economic data (to analyze the sustainability of our products and processes). In the next table the times needed to obtain a valid answer depending on the different datasets is show. The time is in hours.

Economic data	Social data	Environmental data
1,44	0,76	5,54
0,46	3,53	5,59
2,05	2,04	6,36
3,19	1,89	5,38
3,61	0,98	4,5
2,16	1,46	6,76
2,08	0,98	8,12
1,59	1,08	7,22

Do you think that exist an evidence that the algorithm behaves better depending on the nature of the data? Justify your answers and note what are the hypotheses to be tested to conclude.

	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.17
31	4.16	3.31	2.91	2.68	2.52	2.41	2.32	2.26	2.20	2.15
32	4.15	3.30	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.29	2.89	2.66	2.50	2.39	2.30	2.24	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11

3 Problem (2.5 points): Design of Experiments

A bank is planning to reshape its ATM machines. There is space for 6 machines and, we know that we have 3 types, GENERAL ATM, INCOME ATM and MANAGEMENT ATM We conduct a simulation experiment following a 2^3 factorial design and we obtain the next table.

	GENERAL ATM. (A)	INCOME ATM. (B)	MANAGEMENT ATM. (C)	ANSWER
E1	-	-	-	8.7
E2	-	-	+	8.7
E3	-	+	-	8.7
E4	-	+	+	8.7
E5	+	-	-	1.4
E6	+	-	+	1.4
E7	+	+	-	1.4
E8	+	+	+	1.4

Where "+" means 2 servers while "-" means one.
Using the main effects as criteria element, where we are going to invest? (justify your answer).
Develop the calculus using (i) Yates algorithm and (ii) direct formulas.

4 Problem (2.5 points): simulate by hand

Simulate an **activity scanning** engine, suppose a delta = 1. We simulate until time = 6. The priority is for the exits.

Id	Time	Event time	Next Arrival	Next Exit	Server state	Queue long
0	0	0	1	-	0	0

The table that defines the arrival time and the service time is.

Element	Arrival time	Service time
1	1	1.5
2	2	2
3	2.5	3
4	3	1
5	6	2

5 Answer 1

First, we write the table.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Then

X	Y	Y'	Y-Y'	(Y-Y')2
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

The formula for a regression line is

- Y' = bX + A

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.
The equation for the line in Figure is

- **Y' = 0.425X + 0.785**

6 Answer 2

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Economic data	8	16,58	2,0725	0,979136
Social data	8	12,72	1,59	0,823743
Environmental data	8	49,47	6,18375	1,347713

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	101,9673	**2**	50,98363	**48,54673**	**1,33E-08**	**3,4668**
Within Groups	22,05414	**21**	1,050197			
Total	124,0214	23				

We see a huge evidence that suggest that the simples belong to different populations (p-value is clearly small that 0.05, we can conclude that the different samples belong to different populations).
We REJECT H0.
We need to test **homoscedasticity**, **normality** and **independence** of the data to assume the conclusion of the ANOVA model.

7 Answer 3

2.1 To calculate the interactions of A and B:

$$Interactions_{A-B} = \frac{R_1 + R_2 - R_3 - R_4 - R_5 - R_6 + R_7 + R_8}{4}$$

One can use also the Yates algorithm. Both approaches are correct.

2.2

$$Main\_Effect_A = \frac{(E_5 - E_1) + (E_6 - E_2) + (E_7 - E_3) + (E_8 - E_4)}{4}$$
$$= \frac{(1.4 - 8.7) + (1.4 - 8.7) + (1.4 - 8.7) + (1.4 - 8.7)}{4} = \frac{-29.2}{4} = -7.3$$
$$Main\_Effect_B = \frac{(E_3 - E_1) + (E_4 - E_2) + (E_7 - E_5) + (E_8 - E_6)}{4}$$
$$= \frac{(8.7 - 8.7) + (8.7 - 8.7) + (1.4 - 1.4) + (1.4 - 1.4)}{4} = \frac{0}{4} = 0$$
$$Main\_Effect_C = \frac{(E_2 - E_1) + (E_4 - E_3) + (E_6 - E_5) + (E_8 - E_7)}{4}$$
$$= \frac{(8.7 - 8.7) + (8.7 - 8.7) + (1.4 - 1.4) + (1.4 - 1.4)}{4} = \frac{0}{4} = 0$$

Again, one can use Yates algorithm to obtain the values.

A	B	C	ANSWER				
-	-	-	8,7	17,4	34,8	40,4	5,05
-	-	+	8,7	17,4	5,6	0	0
-	+	-	8,7	2,8	0	0	0
-	+	+	8,7	2,8	0	0	0
+	-	-	1,4	0	0	-29,2	-7,3
+	-	+	1,4	0	0	0	0
+	+	-	1,4	0	0	0	0
+	+	+	1,4	0	0	0	0

Factor (A) GENERAL ATM, is the only one that is really important to improve the behavior of the system. The recommendation is to improve this factor. One can detect this just with a simple inspection of the table. Notice that only when one modify factor A the values of the ANSWER are modified.
Note: improving depends on the meaning; as an example, if the answer is the time needed to serve a client, less is better. In that case, scenarios E5 to E8 will be the bests alternatives. Since we do not have more information regarding other factors, like the cost of the servers, we cannot detail nothing more.

8 Answer 4

We put in parenthesis, the time when the event will be really processed due to the Activity Scanning approach.

Id	Time	Event time	Next Arrival	Next Exit	Server state	Queue long
0	0	0	1	-	0	0
1	1	1	2	2.5 (3)	1	0
2	2	2	2.5 (3)	2.5 (3)	1	1
3	3	2.5	2.5 (3)	4.5 (5)	1	0
4	3	2.5	3	4.5 (5)	1	1
5	3	3	6	4.5 (5)	1	2
6	4					
7	5	4.5	6	7.5 (8)	1	1
8	6	6	-	7.5 (8)	1	2