**MASTER IN DATA SCIENCE (FIB-UPC).**

**ACADEMIC YEAR 22-23 Q1** – **PARTIAL EXAM**

**Statistical Inference and Modelling (SIM) .**

**Date: Extraordinary/2022 18:30-20:30 h Classrooms TBD**

**Professor**: Lídia Montero and Josep Franquet

**Rules for quiz:** Internet access is required, emailing and chatting is strictly forbidden. Mobile phones should be switched off. R document folder on the ATENEA are allowed during the exam

**Duration:** 2h 00 min

**Marks**: Before 14/Nov/22 Subject ATENEA WEB site.

**Open Office**: Email requests.

**Problem 1: All qüestions account for 1 point (choose 10 out of 14)**

Cross-section data from the High School and Beyond survey conducted by the Department of Education in 1980, with a follow-up in 1986. The survey included students from approximately 1,100 high schools. Rouse (1995) computed years of education by assigning 12 years to all members of the senior class. Each additional year of secondary education counted as a one year. Students with vocational degrees were assigned 13 years, AA degrees were assigned 14 years, BA degrees were assigned 16 years, those with some graduate education were assigned 17 years, and those with a graduate degree were assigned 18 years. Data frame is available with AER Package. Data frame contains 4,739 observations on 14 variables. **Use df data frame.**

|  |  |  |
| --- | --- | --- |
| 1 | gender | factor indicating gender. |
| 2 | ethnicity | factor indicating ethnicity (African-American, Hispanic or other). |
| 3 | score | base year composite test score. These are achievement tests given to high school seniors in the sample. |
| 4 | fcollege | factor. Is the father a college graduate? |
| 5 | mcollege | factor. Is the mother a college graduate? |
| 6 | home | factor. Does the family own their home? |
| 7 | urban | factor. Is the school in an urban area? |
| 8 | unemp | county unemployment rate in 1980. |
| 9 | wage | state hourly wage in manufacturing in 1980. |
| 10 | distance | distance from 4-year college (in 10 miles). |
| 11 | tuition | average state 4-year college tuition (in 1000 USD). |
| 12 | education | number of years of education. |
| 13 | income | factor. Is the family income above USD 25,000 per year? |
| 14 | region | factor indicating region (West or other). |

***SOURCE:***

* Online complements to Stock and Watson (2007). Usage data("CollegeDistance")

***References***

* Rouse, C.E. (1995). Democratization or Diversion? The Effect of Community Colleges on Educational Attainment. Journal of Business \& Economic Statistics, 12, 217–224.
* Stock, J.H. and Watson, M.W. (2007). Introduction to Econometrics, 2nd ed. Boston: Addison Wesley.

**Distance to school (distance) is going to be our numeric target and income our target factor when needed.**

**> summary(df)**

**gender ethnicity score fcollege mcollege home urban unemp**

**male :2139 other :3050 Min. :28.95 no :3753 no :4088 no : 852 no :3635 Min. : 1.400**

**female:2600 afam : 786 1st Qu.:43.92 yes: 986 yes: 651 yes:3887 yes:1104 1st Qu.: 5.900**

**hispanic: 903 Median :51.19 Median : 7.100**

**Mean :50.89 Mean : 7.597**

**3rd Qu.:57.77 3rd Qu.: 8.900**

**Max. :72.81 Max. :24.900**

**wage distance tuition education income region**

**Min. : 6.590 Min. : 0.000 Min. :0.2575 Min. :12.00 low :3374 other:3796**

**1st Qu.: 8.850 1st Qu.: 0.400 1st Qu.:0.4850 1st Qu.:12.00 high:1365 west : 943**

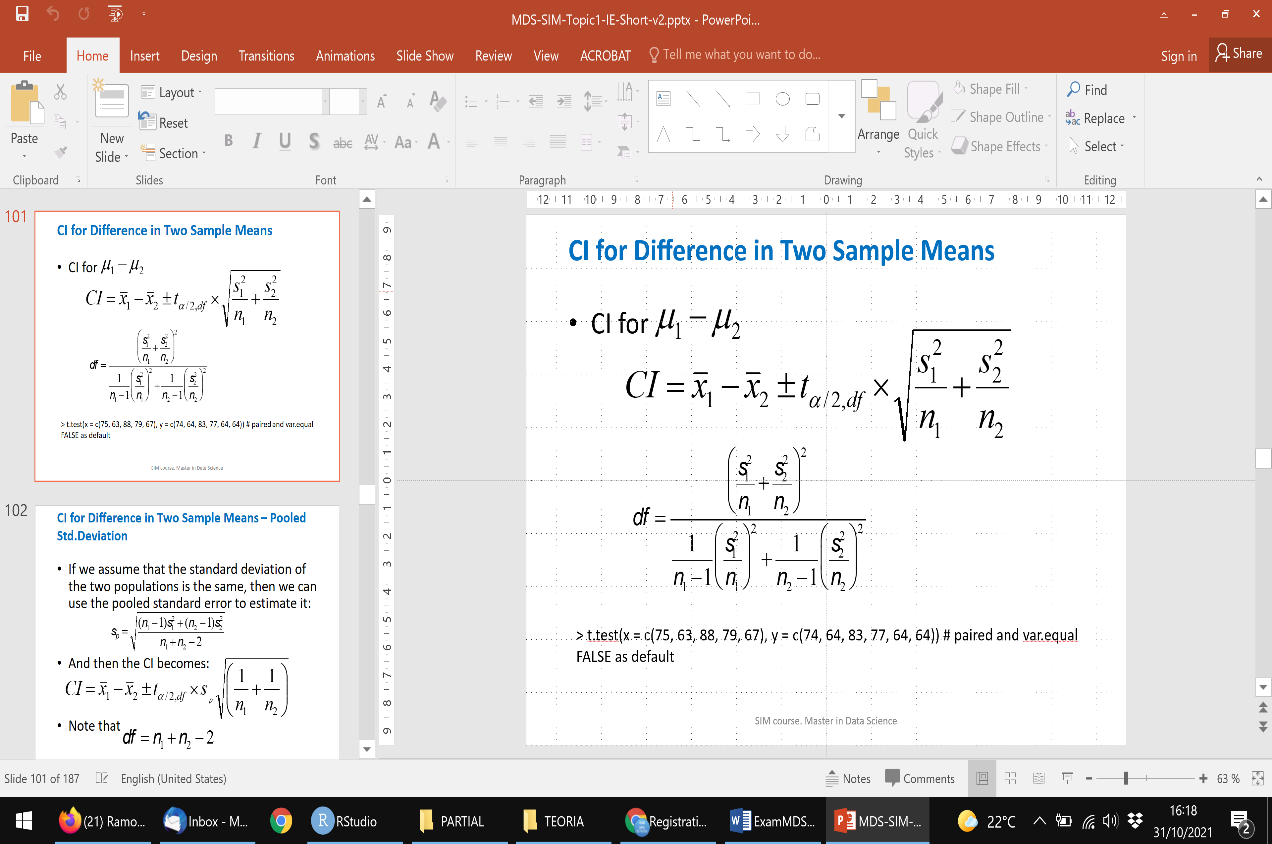
**Median : 9.680 Median : 1.000 Median :0.8245 Median :13.00**

**Mean : 9.501 Mean : 1.803 Mean :0.8146 Mean :13.81**

**3rd Qu.:10.150 3rd Qu.: 2.500 3rd Qu.:1.1270 3rd Qu.:16.00**

**Max. :12.960 Max. :20.000 Max. :1.4042 Max. :18.00**

1. Determine thresholds for mild and severe outliers for wage. Are there any outliers? Indicate the number of mild and severe outliers.
2. Replace by NA those mild outliers in wage variable detected in Point 1 and use an imputation procedure discussed in class to fill outlier data points. Assess the consistency of imputed value/s. ***Remove from dataset those observations with NA in wage variable for the rest of questions.***
3. Would you expect a family paying tuition fees of 900$ to have a shorter or longer college distance than a family paying 270$?
4. Analyse the profile of the numeric target (distance) using condes() method. A detailed explanation of procedure results is requested.
5. Analyse the profile of the categorical target (income) using a suitable method. A detailed explanation of procedure results is requested when profiling low income (high income may be omitted).
6. Discuss whether a normal distribution would be a reasonable distribution for distance target.
7. Is there variance homogeneity in the distance target groups defined by ethnicity levels? Assess race characteristics.
8. Distance target can be considered to be the equal across groups defined by by ethnicity levels? Use a two.sided test at 99% confidence.
9. State and test one.sided hypothesis to assess whether distance is greater for the afroamerican group than for the Hispanic group or the opposite at 99% confidence.
10. The standard deviation of distance in afam ethnicity group should not exceed 2. For the sample in f.hcla 1 in your dataset, calculate the deviation of distance assuming a normal distribution. Stating any assumptions, you need (write them), test at the 1% level the hypothesis that the population standard deviation is larger than 10,000$.
11. Figure out the 99% upper threshold for distance in afam ethnicity subpopulation variance. Normal distribution for distance is assumed to hold.
12. Build a 99% two-sided confidence interval for the difference in the mean of distance between Hispanic and Afroamerican ethnicity groups. Assume that equal variances in the population distance does not hold and normal distribution of distance (to simplify the calculations) does hold, but justify if these assumptions are critical.
13. Determine a 99% confidence interval for the population proportion that represents a low income. Test the null hypothesis that low and high income groups have equal probability.
14. A new survey considered 1000 people, 660 were classified in the low income group. Determine a 99% confidence interval for the difference in the population proportion of low income accounting for the two sources. Test the null hypothesis that having a low income has a lower incidence in the survey than in the original sample.

Hint:

