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

**ACADEMIC YEAR 23-24 Q1** – **MIDTERM EXAM**

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

**Date: 3/Nov/2023 15:30-17:30 h Classrooms A5S102 - Group 11, A5S111-Group 12**

**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 documents in folder on the ATENEA Virtual Campus are allowed during the exam

**Duration:** 2h 00 min

**Marks**: Before 14/Nov/23 Subject ATENEA WEB site.

**Open Office**: November 15th 2023 at 20h after class. Email requests also allowed.

**Problem 1: All questions account for 1 point (you have to answer all of them)**

Dataset for this exercise can be found on stata website (<http://www.stata.com/texts/eacsap/>) and it was used to produce the output in the book ***[Econometric Analysis of Cross Section and Panel Data](https://www.stata.com/bookstore/econometric-analysis-cross-section-panel-data/)*** by Jeffrey Wooldridge. There are 680 observations about class attendance and results on assignments.

-----------------------------------------------------------------------------

storage display value

variable name type format label variable label

-----------------------------------------------------------------------------

attend byte %8.0g classes attended out of 32

termgpa float %9.0g GPA for term

priGPA float %9.0g cumulative GPA prior to term

ACT byte %8.0g ACT score

final byte %8.0g final exam score

atndrte float %9.0g percent classes attended

hwrte float %9.0g percent homework turned in

frosh byte %8.0g =1 if freshman

soph byte %8.0g =1 if sophomore

skipped byte %9.0g number of classes skipped

stndfnl float %9.0g (final - mean)/sd (target)

A descriptive statistic summary() based on R shows:

> summary(attend)

attend termgpa priGPA ACT final

Min. : 2.00 Min. :0.000 Min. :0.857 Min. :13.00 Min. :10.00

1st Qu.:24.00 1st Qu.:2.138 1st Qu.:2.190 1st Qu.:20.00 1st Qu.:22.00

Median :28.00 Median :2.670 Median :2.560 Median :22.00 Median :26.00

Mean :26.15 Mean :2.601 Mean :2.587 Mean :22.51 Mean :25.89

3rd Qu.:30.00 3rd Qu.:3.120 3rd Qu.:2.943 3rd Qu.:25.00 3rd Qu.:29.00

Max. :32.00 Max. :4.000 Max. :3.930 Max. :32.00 Max. :39.00

atndrte hwrte frosh soph skipped

Min. : 6.25 Min. : 12.5 Min. :0.0000 Min. :0.0000 Min. : 0.000

1st Qu.: 75.00 1st Qu.: 87.5 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 2.000

Median : 87.50 Median :100.0 Median :0.0000 Median :1.0000 Median : 4.000

Mean : 81.71 Mean : 87.9 Mean :0.2324 Mean :0.5765 Mean : 5.853

3rd Qu.: 93.75 3rd Qu.:100.0 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.: 8.000

Max. :100.00 Max. :100.0 Max. :1.0000 Max. :1.0000 Max. :30.000

NA's : 6.0

stndfnl

Min. :-3.30882

1st Qu.:-0.78782

Median : 0.05252

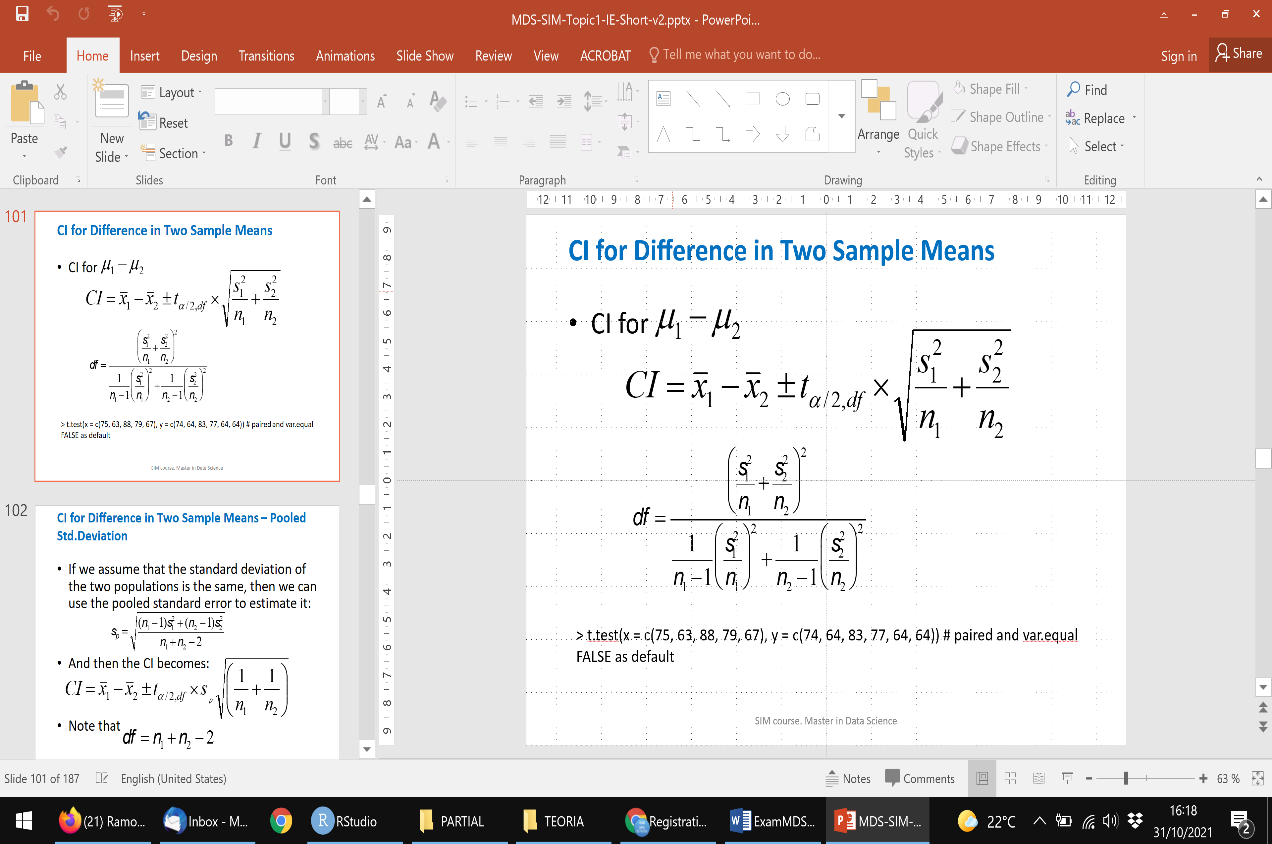
Mean : 0.02966

3rd Qu.: 0.68277

Max. : 2.78361

**Selected numeric target is stndfnl (standardized final exam score) and let freshman status be the qualitative target. Prepare your dataset to represent factors in a suitable way.**

1. Determine thresholds for mild and severe outliers in the target. Are there any outliers? Indicate observation id’s and atypical values. ***Do not take any action****.*
2. Use an imputation method to address missing data in the dataset. Validate imputation results.
3. Are there any multivariate outliers in the dataset? Indicate how many and which at 99% confidence. Explain what they seem to share in common. ***Do not take any action****.*
4. Analyze the profile of the numeric target (**stndfnl**) using a suitable profiling method. A detailed explanation of the procedure outcome is requested.
5. Analyze the profile of the binary target (**frosh**) using a suitable method. A detailed explanation of the procedure outcome is requested.
6. Is there variance homogeneity in the **stndfnl** target groups defined by frosh and soph classes, one by one and when applied simultaneously? *Hint: You have to define a new factor.*
7. Mean **stndfnl** target can be considered to be the equal across groups defined by frosh target? Use a two.sided test at 1% significance level and indicate the confidence interval for freshman target population mean. Indicate whether equal variances and normal distribution of stndfnl hypothesis hold in the population.
8. State and test one.sided hypothesis to assess whether **stndfnl** is less for freshman than the rest at 1% significance level. Indicate and justify a 95% confidence interval for freshman target population mean.
9. The standard deviation of **stndfnl** in the freshman group should not exceed 1. For the sample in the freshman group in your dataset, calculate the deviation of **stndfnl** assuming that normal assumption holds. State and include any assumptions needed to test at the 1% significance level whether population standard deviation is larger than 1 in the freshman group. Figure out the 99% upper threshold for **stndfnl** in the freshman population standard deviation.
10. Determine a 99% confidence interval for the population proportion of neither freshman, nor sophomore. Test the null hypothesis that selecting a unit neither freshman, nor sophomore has a population probability greater than 0.25.
11. Test the null hypothesis that the proportion of freshman group and the population proportion of neither freshman, nor sophomore is the same at 1% significance level.

Hint:

