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

**ACADEMIC YEAR 21-22 Q1** – **PARTIAL EXAM**

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

**Date: 8/Nov/2021 16:00-18:00 h Classrooms C6S301 C6S302**

**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 17/11/21 Subject ATENEA WEB site.

**Open Office**: Email requests.

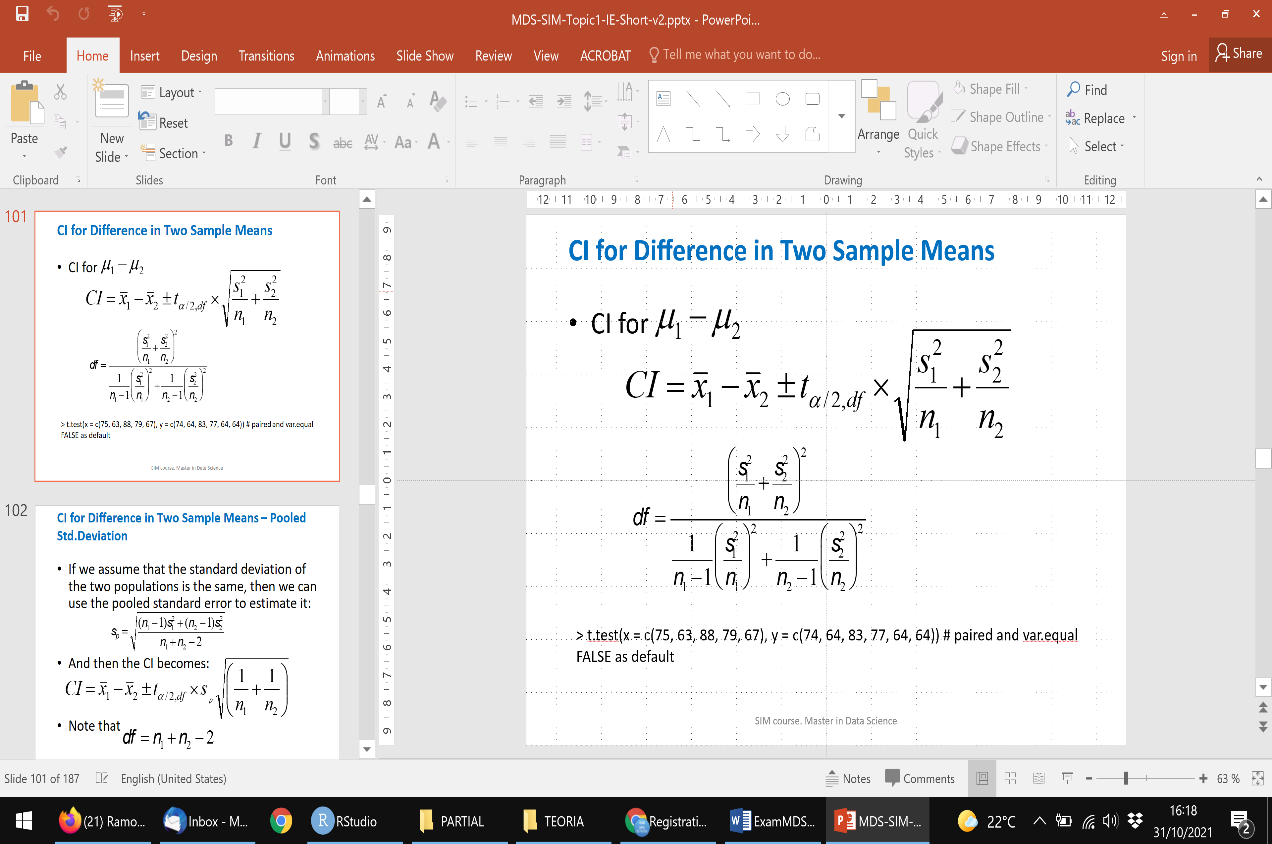
**Problem 1: All questions account for 1 point**

**Nutrient analysis of pizzas** (from https://data.world/sdhilip/pizza-datasets). Who likes pizza? I mean, there are so many things to like, let’s take a closer look! The data set pizza. Pizza.RData contains measurements that capture the kind of things that make a pizza tasty. The variables in the data set are:

|  |  |
| --- | --- |
| Feature | Description |
| brand | Pizza brand (class label) – Target factor |
| id | Sample analysed |
| mois | Amount of water per 100 grams in the sample |
| prot | Amount of protein per 100 grams in the sample |
| fat | Amount of fat per 100 grams in the sample |
| ash | Amount of ash per 100 grams in the sample |
| sodium | Amount of sodium per 100 grams in the sample |
| carb | Amount of carbohydrates per 100 grams in the sample |
| cal | Amount of calories per 100 grams in the sample – Numeric Target |

**The amount of calories per 100 grams is selected to be the numeric target. In the past, a lognormal distribution was accepted as the reference distribution.**

1. Determine whether serial correlation on the amount of calories is present on dataset or not.
2. Define a new variable containing the total amount of ingredients (water, protein, fat, ash and carbohydrates). Check consistency.
3. Univariate severe outliers are also present in some variables. Summarize the total number of severe outliers.
4. Are there multivariate outliers? Find them. Try to explain their singularity. Multivariate outliers are not going to be treated in this exercise: keep them as supplementary observations. Which is the cut-off at 99.9% CI?
5. Indicate by using exploratory data analysis tools which are apparently the most associated variables with the numeric response variable (only the contributing variables to the ingredients are to be taken as active variables). Use also **FactoMineR profiling tools** at 99% significance level.
6. Use brand target factor and determine the most remarkable global associations at 99% CI. Profile A and I brands.
7. Say a few words about the hypothetical distribution that was assumed in the past: is it reasonable?. Use graphical and inferential arguments at 95% confidence.
8. Let us focus on **cal** variatedispersion behavior according to the brand. Use numeric, graphics and inferential tools at 95% confidence to address the topic.
9. Let us focus on **cal** variatemean behavior according to the brand. Use numeric, graphics and inferential tools at 95% confidence to address whether the mean of the target depends on the brand or not.
10. Continuing with the former question, on the positive case which brands show a remarkable difference in mean behavior among them at 95% confidence. Use one-sided tests to address the question.
11. The standard deviation of the number of calories for brand A should not exceed 0.15cal. For the simple random sample in your dataset, calculate the deviation of calories for 100g assuming a normal distribution for 100 g calories. Stating any assumptions, you need (write them), test at the 1% level the null hypothesis that the population standard deviation is not larger than 0.15cal against the alternative that it is
12. Figure out the 99% upper threshold for the number of calories for brand A population variance. Normal distribution for calories is assumed to hold.
13. Build a 99% confidence interval for the difference in the mean of 100 g calories between brands A and C. Assume that equal variances in the population calories per brand does not hold and normal distribution of calories per 100g (to simplify the calculations), but justify if these assumptions are critical.
14. Out of 100 people, 60 prefer A to C. Determine a 99% confidence interval for the population proportion that favors A in front of C. Test the null hypothesis that selecting A and C has equal probability.
15. A second survey considered 200 people, 110 prefer A to C. Determine a 99% confidence interval for the difference in the population proportion that favors A in front of C accounting the two surveys. Test the null hypothesis that selecting A brand has a lower probability in the second of the surveys.

Hint:

