STATISTICAL MODELS

Course 2024-2025

LAB 2. Multivariate Statistics

PROBLEM 1.- File `Head_Brothers.xlsx` contains measurements of the length and breadth of the heads of pairs of adult brothers in 40 randomly sampled families. We wish to analyze this dataset with techniques that assume multivariate Gaussianity:

a) Assess whether each of the 4 dimensions of these data is normally distributed. Use $\alpha = 0.01$ as significance level.

b) Assess whether each pair of dimensions is normally distributed.

c) Compute a chi-squared plot for all 4 dimensions considered together.

d) Identify possible outliers. If one or more outliers are found, concisely discuss what shall be done with it/them.

e) Based on all the above information, indicate whether you would find it acceptable to test this dataset using statistical techniques that assume the data following an (approximately) multivariate normal distribution. Make a clear and concise justification of your answer.

PROBLEM 2.- A group of 100 subjects is administered a new treatment and we are asked to analyze whether it produces a significant effect. To this end, we are provided with two random samples:

- File `Lab_2_Before.xlsx` contains measurements for the overall physical and mental health (two values per subject, in a continuous scale from 0=poor health to 100=excellent health) before the treatment.
- File `Lab_2_After.xlsx` contains measurements for the overall physical and mental health (two values per subject, in a continuous scale from 0=poor health to 100=excellent health) after the treatment.

a) State the null and alternative hypotheses for this problem.

b) Indicate what method would you use to answer the research questions of this Problem.

c) State the assumptions required by your method and assess whether the provided data fulfils all of them. In case assumptions are not met, go back to (b) and choose a different method, repeating (c) until all assumptions are fulfilled.

d) Apply the method selected in (b) to answer the research question, at significance level $\alpha = 0.05$. Indicate your conclusions.

e) Construct a 95% confidence region for the (population) mean difference in health measurements before and after treatment. Analyze whether the resulting region confirms your conclusion in (d).

PROBLEM 3.- In this problem we will analyze continuous evaluation data that corresponds to Statistical Models from 2019 to 2021. To this end, file `SM_SamplesMarks_2019_2021.xlsx` contains data from 78 students. For each of them (corresponding to a row in the referenced file), we have collected:

1. The year in which the student took the course.
2. The marks obtained in each of the 5 labs with deliverables.
3. The resulting Continuous Evaluation mark (which is simply the average of the lab marks).
4. The result of the student in the final exam, as a binary variable ($1 =$ pass).

Students included in the file were selected randomly according to the following criteria:

- For each year, 13 students who passed the exam and 13 students who did not pass de exam were selected.
- Students with a Continuous Evaluation mark below 5.0 were not eligible, since they could not take the final exam.

We wish to assess whether, on average, there is a difference considering all the 5 Lab Marks of students (not their average), depending on:

- Whether they passed the final exam or not.
- The year in which they took the course.

Using significance level $\alpha = 0.01$, answer the following statements:

a) Indicate what method would you use to answer the research questions of this Problem. Indicate the assumptions that should be fulfilled (make sure to specify them clearly and adjusted to the data to be used; e.g. statements like "independence" without further clarification will not be accepted). For this problem, you do not need to test whether the assumptions hold; we will assume that they do.

b) Test whether there is a significant interaction effect. Clearly state: $i)$ the corresponding null and alternative hypotheses, $ii)$ all the relevant numerical details, and $iii)$ your conclusion.

c) Based on your results from (b), indicate whether further analysis of these data should be based on assessing *main effects* or *simple effects*. Justify adequately, including any additional analysis that you believe necessary to provide an adequate answer.

d) Proceed to analyze main[1] or simple effects, according to your answer in (b). In all cases provide your conclusions, clearly indicating what they imply for the two research questions proposed in this Problem.

[1] In case of analyzing main effects, make sure to provide a complete analysis, e.g. including any post-hoc tests that might be necessary to determine what groups differ, if any.