STATISTICAL MODELS
Course 2024-2025

LAB 4. Resampling

PROBLEM 1.- File `SM23_LabMarks.xlsx` shows the lab marks obtained by the 74 matriculated students in Statistical Models in course 2023-2024. Similarly to the present course, labs 1, 3 and 5 consisted on small exams, while labs 2, 4 and 6 consisted of deliverables. We with to use the data from those 74 students as a random sample to make inferences about the continuous evaluation method that we are using.

More specifically, we wish to assess whether the tests are more difficult than the deliverables, for which we wish to make inferences about the following parameter:

$$\theta = E[\eta] \qquad \eta = \frac{Average\_mark\_Exams}{Average\_mark\_deliverables}$$

where $E[]$ stands for the expected value[1]. In other words, for each student we are interested in the ratio between his/her average marks on the exams and his/her average marks on the deliverables, which we call η, and we set the expected value of these ratios as the parameter of the population to be analysed to answer our research question. We will use $\alpha = 0.05$ as the significance level.

a) State the null and alternative hypotheses for this problem, including the mathematical formulation in terms of the requested parameter.

b) Define a suitable statistic to make inferences about the parameter of interest. Compute this statistic on the provided samples and analyse the obtained values looking for any outliers or invalid data. Be careful with the assumptions that you use (since we expect to use bootstrap in this lab there is no need for normality); justify very well the reasons that allow the exclusion of any data sample in case you decide that this is needed.

c) Use the sample values computed in (b) to assess whether η is normally distributed.

d) Theory suggests that, given the relatively large number of samples, the statistic that you selected in (b) should be normally distributed, regardless of your answer in (c)[2]. Use resampling to confirm whether this holds.

e) Construct an adequate confidence interval for the parameter of interest, in such a way that you can use it to answer our research question. Provide the answer and a clear justification for it.

f) Use resampling to test the hypotheses in (a) directly and provide the obtained p-value; compare the result to the one obtained in (e).

[1] Be aware that this expectation would correspond to the population of all possible students taking the subject under this evaluation methodology, e.g. all UPF engineering students.

[2] Reflect about this and, if you find that it is not the case, then you have most likely chosen a wrong statistic in (b)

PROBLEM 2.- Using the same data used in Problem 1, we wish to address now a different question: are the exam and deliverable marks of students independent? In other words is there a clear relation between the (average) exam and deliverable marks obtained by each student, or they are simply independent because exams and deliverables are so different ways of assessing students?

Again, we will use $\alpha = 0.05$ as the significance level, and now the parameter of interest will be the correlation coefficient between the average-exam and average-lab marks.

a) State the null and alternative hypotheses for this problem, including the mathematical formulation in terms of the requested parameter.

b) Using the sample correlation as your statistic, compute its values using only the samples that you decided to be valid in Problem 1(b). Use resampling to assess whether the chosen statistic follow a normal distribution.

c) Construct an adequate confidence interval for the parameter of interest, in such a way that you can use it to answer our research question. Provide the answer and a clear justification for it.

f) Use resampling to test the hypotheses in (a) directly and provide the obtained p-value; compare the result to the one obtained in (c).

PROBLEM 3.- To design clothing, a company selling excercise clothing has collected data on the physical characteristics of customers. File `SM_Lab_4_WeightsRunners.xlsx` provides the weights (kg) for a random sample of 25 male runners. Since products target the *average* male you are interested in seeing how much the subjects in your sample vary from the average weight.

a) Calculate the sample standard deviation $\hat{\sigma}$ for these weights and find the bias of $\hat{\sigma}$.

b) Use bootstrapping to find the bootstrap standard error for $\hat{\sigma}$.

c) Make a histogram of the bootstrap replications of $\hat{\sigma}$ from (b). Inspect the data to try assessing why do we obtain such a *strange* histogram.

d) Suggest a way to improve the analysis so that you can compute a confidence interval using bootstrapping. Clearly justify why your proposal is adequate.

e) Implement the solution proposed in (d) and compute a 95% confidence interval for $\hat{\sigma}$. Clearly indicate what method you choose to compute the interval.