Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 7th, 2020

Problem n.1

The file `weather.txt` reports weather data collected on the first months of the year 2020 in Milan. Each statistical unit corresponds to a day. For each unit, the following measurements are reported: mean temperature (°C), min temperature (°C), max temperature (°C), dew point (°C), humidity (%), visibility (km), mean wind (km/h) and max wind (km/h).

a) Perform a Principal Component Analysis of the dataset, evaluate whether it is appropriate to use the original variables or the standardized ones and proceed accordingly. Report a plot of the loadings of the first 3 principal components and interpret them.

b) Report the scatter plot of the data along the first two principal components and describe how to interpret the four quadrants of the plane. Use the information about the day of the year to further interpret the results.

c) Report the screeplot. Propose (with motivations) a dimensionality reduction of the dataset. Report the variance explained along the principal components retained.

d) The measurements for the 1^{st} of August are the following: mean temperature 30°C, min temperature 23°C, max temperature 36°C, dew point 22°C, humidity 65%, visibility 19km, mean wind 5km/h and max wind 15km/h. Project the new datum on the reduced space identified at point (c).

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUNElYQjFTNkRFUDE4UExUUjlGSk4wV0VaRi4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 7th, 2020

Problem n.2

To compare the performances of the led bulbs produced by two brands (Candle Inc. and Sunshine Corp.), brightness tests are performed on 50 led bulbs independently sampled from those produced by Candle Inc. and 50 led bulbs independently sampled from those produced by Sunshine Corp.

For each led bulb, the brightness is measured by two light meters positioned at different distances from the led bulb: the first light meter is a few centimeters from the led bulb and the second is positioned at one meter from the led bulb. The measurements on different led bulbs are independent (while the measurements on the same led bulb are not).

Files `candle.txt` and `sunshine.txt` report the measurements obtained with the two light meters on each led bulb produced by Candle Inc. and Sunshine Corp. respectively.

a) Perform a statistical test of level 95% to verify if the mean measurements on the led bulbs produced by the two brands differ. State the model assumptions required to perform the test and verify them.

b) Compute and report the p-value of the test at point (a).

c) Interpret the results of the test at point (a) through two Bonferroni intervals of global level 95% for appropriate differences in the mean. Comment the result.

d) Is there statistical evidence to state that, at level 95%, the decrease in brightness between the two measurements of the led bulbs produced by Candle Inc. is in mean higher than the one of the led bulbs produced by Sunshine Corp.?

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUQVZBMzlBMUdQSzZZTFZUOFVPOEY1TU5GOS4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 7th, 2020

Problem n.3

Benedetta is learning to bake bread. To obtain the perfect dough rising, she is performing some experiments to evaluate the effect of the kind of yeast used and the effect of the waiting time (measured in hours) on the rising of the dough. In the different experiments, she used either the brewer's yeast (`by`) or the sourdough (`sd`) and measured the rising of the dough as a percentage with respect to the initial volume at different times. The file `leaven.txt` contains the data collected by Benedetta in 80 experiments conducted in the last months.

Consider a linear model assuming, for each kind of yeast, a polynomial of degree at most 2 in waiting time (`time`). Impose the same intercept for the two models.

a) Estimate the parameters of the model. Verify the model assumptions.

b) Perform two statistical tests - each at level 1% - to verify if

   - there is a significant dependence of the rising of the dough on the kind of yeast;
   - there is statistical evidence that the degree of the polynomial for the brewer's yeast is higher than the degree of the polynomial for the sourdough.

c) Based on the tests performed in (b) or any other test deemed relevant, reduce the model and update the model parameters.

d) Benedetta wants to bake bread for dinner, but she can wait only 2 hours: which yeast do you suggest to use? Provide a confidence interval with confidence level 99% for the mean of the rising of the dough in such conditions.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUNlJSRUxSV0JZNDNESTc1WFdFSFdETFdCQS4u`

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 7th, 2020

Problem n.4

To forecast the tidal currents in the port of Dublin, the vertical motion induced by the rise and fall of the tides is monitored. The file `tide.txt` reports the measurements of the sea level in the port of Dublin collected every half hour for one day. Since the time variation of the sea level is inherently a continuous process, consider a functional data analysis approach. It is known that the available measurements are affected by small errors.

a) Perform a smoothing of the data using a B-spline basis of degree 3. Choose the number of basis functions using a generalized cross-validation criterion. Report the number of basis functions chosen, a plot of the B-spline basis system used and a plot of the smoothed data.

b) Compute approximate pointwise confidence intervals at the sampling times of the data and provide a plot.

c) To study the currents induced by the tide, it is important to have a good estimate of the velocity of the tide. Compute an approximation of the first derivative of the curve from the data and the first derivative of the smoothed curve obtained at point (a). Provide a plot to compare the two and comment on the result.

d) Would you suggest the use of a different basis system? Why?

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUMFRDUVpBMUVNT1BEUUlZNU5BNUo3SjQ4QS4u`