Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
September 13th, 2018

First Name and Family Name:
ID Number:

Problem n.1

The file `IAMG.txt` collects the data regarding the participation to the annual meetings of the International Association for Mathematical Geosciences (IAMG), in the last 50 years. For each meeting, it reports the number of registered participants, the number of oral presentations and the number of no-show (i.e., the number of registered participants that did not show up). Call X the vector whose components are the number of registered participants (X_1), of oral presentations (X_2) and of no-show (X_3), and assume each meeting to be independent of the others.

a) Build a confidence region (level 95%) for the mean of X. Characterize the region by reporting its expression, its center, the direction of the axes and the length of the semi-axes.

b) Build three T^2-simultaneous confidence intervals (level 95%) for: the mean number of registered participants, the mean number of oral presentations and the mean number of no-show.

c) Perform a test of level 95% to verify the hypothesis according to which, in mean, only 90% of the registered participants actually show up at IAMG meetings.

Problem n.2

The file `Waiting.txt` reports the waiting times for food – i.e., the time between the order of a course and the service – in 180 restaurants in Romania. The dataset also reports the type of course (starter, main course or dessert) and the location of the restaurant (Iasi or Bucarest).

a) Propose a complete ANOVA model for the waiting time as a function of the factors *course* (starters, main course or dessert) and *city* (Iasi or Bucarest). Report and verify the assumptions of the model.

b) Comment on the significance of the factors and of their interaction. If needed, propose a reduced model.

c) Build Bonferroni confidence intervals (global level 95%) for the mean differences between the waiting times in the groups identified at point (b), and for the variances of the waiting times within the groups. Comment the results.

Problem n.3

The file `Sailing.txt` collects the data on the daily values of consumed water [l/day] and sailing time [min/day] for 120 sailing cruises in Croatia, in August 2018. It also reports whether the sailboat was occupied by expert sailors (*seadog*) or by inexperienced vacationers (*vacationer*). It has been estimated that in August, on average, only 20% of sailboats are occupied by expert sailors.

a) Based on the available features, build two Bayes classifiers, A and B, for the kind of sailboat' occupation (*vacationer, seadog*), by assuming that:

   A. the two populations are Gaussian with the same covariance structure;

   B. the two populations are Gaussian with different covariance structures.

   For each classifier report a qualitative plot of the classification regions, and the estimated posterior probability associated with the first observation (*water = 32.08, sailing.time = 82.69*).

b) Evaluate the performances of the classifiers A and B and identify the best one.

c) How would you classify the occupants of a sailboat with daily consumed water 35 l and daily sailing time 168 min?

Problem n.4

The file `Lumieres.txt` reports the data on the participation to *Rendez-vous*, the show of sounds and lights which takes place every evening between the 1st June and the 31st August in the main square of Nancy (France). The dataset refers to the years 2016 to 2018, and reports the number of participants (n), the day of the representation (d, with $d = 1$ on the 1st June, ..., $d = 92$ on the 31st August), the temperature recorded that evening at 10 pm ($temp$) and the weather conditions ($rain$: yes/no). Consider the following model

$$n_{i,g} = \beta_{0,g} + \beta_1 \cdot d_i + \beta_2 \cdot d_i^2 + \beta_3 \cdot temp + \epsilon,$$

where $g \in \{1, 2\}$ indicates the group according to the weather conditions ($g = 1$ for no rain, $g = 2$ for rain) and $\epsilon \sim N(0, \sigma^2)$.

a) Estimate the 6 parameters of the model. Verify the model assumptions.

b) Perform a statistical test to verify if the mean number of participants depends significantly on the day of the representation.

c) Based on a statistical test of level 95%, reduce the model and update the parameter estimates.

d) Perform a test to verify if the maximum of the expected number of participants is on the last day of July ($d = 61$) and, in case, update the estimates of the model parameters.

e) Based on the last update of the model parameters, provide a prediction interval (probability 95%) for the number of people participating to the representation on the 28th July ($d = 58$, $temp = 29$, no rain).