**Politecnico di Milano**
**Scuola di Ingegneria Industriale e dell'Informazione**
Applied Statistics
June 28th, 2019

***First Name and Family Name:***
***ID Number:***

# Problem n.1

The Catalan Food Association has launched an award for the *Best Catalan Tapa*. As part of the challenge, two tasters are sent to evaluate the 35 finalist tapas in Terrassa and the 35 finalist tapas in Girona. Files `terrassa.txt` and `girona.txt` collect the evaluations on each of the finalist tapas given by the tasters in Terrassa and Girona, respectively. Assume the evaluations on different tapas to be independent, and the evaluations of the two tasters on the same tapa to come from a bivariate Gaussian distribution.

a) Perform a statistical test of level 95% to verify if the mean evaluations in the two cities differ. State and verify the model assumptions.

b) 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.

c) Is there statistical evidence to state that, at level 95%, the *average*$^*$ evaluations of Girona's tapas are in mean higher than those of Terrassa's tapas?

[$^*$by *average evaluation* of a tapa is meant the one obtained by averaging the evaluations of the two tasters on that tapa]

# Problem n.2

The file `profiling.txt` collects the data about the habits of 1067 people in the city of Terrassa, measured on the 4th June 2019 through the smartphone application *FollowYou*. The dataset reports the habits of the users that declared to be either *tourists* or *residents*, in terms of the time $t_1$ [minutes] spent walking in the city centre around Plaza Grande, and the time $t_2$ [minutes] spent waiting for public transportation. The provider asks for your help to build a classifier for profiling new users, in order to provide personalized advertisements.

a) Build a classifier for the variable *type of user* based on the available quantitative features. Report the model for the data, the estimates of its parameters (means and covariances), the priors within the groups and verify the model assumptions. Report a qualitative plot of the classification regions.

b) Compute the APER of the classifier.

c) How would you profile a new user with $t_1 = 35$ min and $t_2 = 3$ min?

# Problema 3

The file `airport.txt` reports the data on the duration of the trips [min] made by 168 travellers moving from Terrassa to Barcellona airport, using the *Express Shuttle Bus* in weekdays of June 2019. For the duration of a trip consider a linear model, accounting for the distance traveled $x$ [km] (i.e., the distance from the bus stop of origin and the airport), and for the time of the day ('6-10', '11-15', '16-20'):

$$Y_g = \beta_{0,g} + \beta_{1,g} \cdot x + \epsilon,$$

with $\epsilon \sim N(0, \sigma^2)$ and $g$ the grouping structure induced by the time of the day.

a) Estimate the parameters of the model ($\{\beta_{0,g}, \beta_{1,g}, \sigma\}$). Verify the model assumptions.

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

   - there is a significant dependence of the mean duration on the time of the day;
   - there is a significant dependence of the mean duration on the distance traveled.

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

d) You have a flight from Barcelona airport at 10:30 a.m., and you want to be at the airport at least 1 hour before the flight departure. At the bus station in front of your hotel in Terrassa (distance of 57 km from the airport), the bus is scheduled to depart every 30 minutes from 6 a.m. to 20:30 p.m.. What time would you take the bus to be on time with probability 99%?

# Problem n.4

The file `montserrat.txt` collects the wind speeds ($w(s_i)$) registered on the 5th June 2019 at 134 measurement sites ($s_i$, $i = 1, ..., 134$), in the region around the Montserrat (Spain). The dataset also reports the UTM coordinates of the measurement sites, and their Euclidean distance to the top of the mountain $d(s_i) = \|s_i - s_0\|$, with $s_0 = (402476, 4605558)$.

a) Estimate two empirical variograms, assuming the following models:
(M1) $w(s_i) = a_0 + \delta(s_i)$;
(M2) $w(s_i) = a_0 + a_1 \cdot d(s_i) + \delta(s_i)$.
Report a qualitative plot of the variograms. Comment the results and choose the model you deem more appropriate for the observations.

b) Fit to the empirical variogram chosen at point (a), a spherical model without nugget, via weighted least squares. Report the estimates of sill and range. Comment the results.

c) Estimate, via Generalized Least Squares, the parameter(s) $a$ of the model chosen at point (a).

d) Predict the wind speed at the top of the mountain, $s_0 = (402476, 4605558)$. Report the associated prediction variance.