ESIGELEC. Course: "Statistics, Probabilities and Optimization for Data Analytics". September 2018.

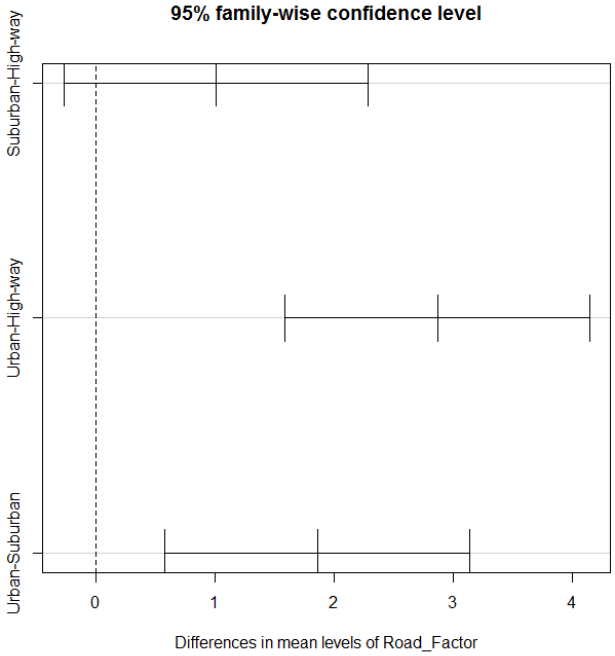
Name: _____ Date: _____

Task 1 (4 points). The ministry of economics of a certain country has conducted a statistical study, about the yearly income per family. This study has shown that the average yearly income per family is 35045 euros. The median of this distribution is 32876, whereas the first and the third quartile are 26769 and 39076 euros, respectively. The standard deviation of the variable studied is 3486 euros. From this information, decide whether the following statements are true, justifying your answer.

a) Half of the families have a yearly income between 26769 and 39076.
b) The outliers are those families whose yearly income is 3486 euros more than the average or 3486 less than the average.
c) 25% of families have an income equal to 26769 or less.
d) 75% of families have an income equal to 39076 euros or more.

Task 2. (6 points) A car company is checking the fuel consumption of a certain car model, of which they sell hundreds of thousands of units along the world. They suspect that this consumption may depend on the type of transmission (manual or automatic) and on the type of journey (highways, suburban areas, and urban areas). For this reason, 24 different cars of this model have been analyzed. For each of these cars, the following characteristics have been measured: "Consumption" shows the fuel consumption after a 100 km road trip, measured in liters; "Transmission" determines the type of transmission of the car (manual or automatic); "Road" determines the types of roads where the trip took place (highways, suburban areas, and urban areas). The car company wants to perform an ANOVA analysis to determine whether the type of transmission and/or the type of road significantly affect the average fuel consumption. The fuel consumption is normally distributed. The ANOVA table and the Tukey HSD intervals for factor "Road" are shown below.

```
                      Df Sum Sq Mean Sq F value   Pr(>F)
Transmission           1   2.46   2.462   2.454   0.1346
Road                   2  33.89  16.944  16.891 7.41e-05 ***
Transmission:Road      2   6.36   3.180   3.170   0.0662 .
Residuals             18  18.06   1.003
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



a) Determine the population, the random variable of interest, the factors and its levels, of this analysis of variance problem.
b) Does the type of transmission significantly affect the fuel consumption? Clearly detail the hypothesis, the p-value, and the conclusion.
c) Does the type of road significantly affect the fuel consumption? Clearly detail the hypothesis, the p-value, and the conclusion.
d) Does "Road" affect "Consumption" differently for different values of "Transmission"? Why?
e) Discuss for which types of road the consumption is expected to be the same, and for which the consumption is expected to be different.

Task 3. (6 points) A human resources department uses a test to hire staff. You know that 60% of applicants pass the test. If an applicant passes the test, they have to go through a training period at the company before being hired. Out of the applicants who pass the test, some 80% of them successfully complete the training period at the company, and are later hired by the company. In order to check the validity of the test, they once also put through the training period the applicants who did not pass the text. They saw that only 50% of these applicants could successfully complete the training period.

a) If no test were used, what percentage of applicants would successfully complete the training period?
b) What is the probability of passing the test for an applicant that has successfully completed the training period?

Task 4. (4 points) The exports of a given country has been analyze by means of a Time Series analysis. The predictions obtained by an ARIMA model over 11 periods are shown below (in red points), as well as its confidence intervals (in red dashed lines). Besides, the actual 11 values of the variable *exports* on these periods is represented as blue points. Discuss on the accuracy of the predictions given. Be as precise as possible and justify your answers.