Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 12th 2019

Surname, name:

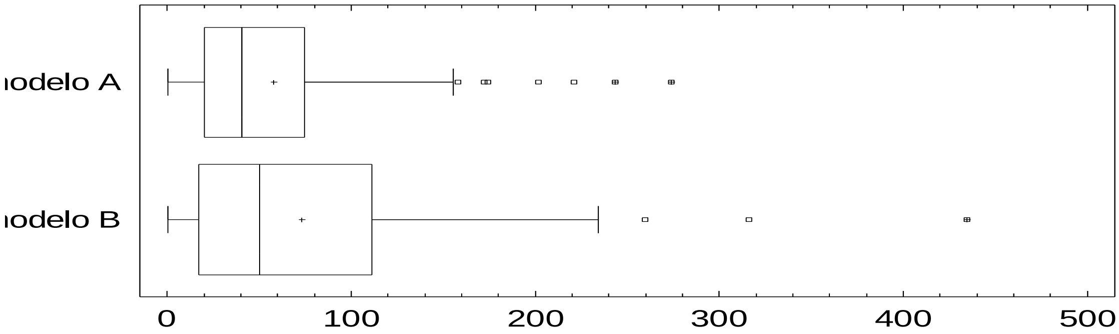
Group: **1E** Signature:

Indicate with a tick mark the partials examined 1st □ 2nd □

Instructions

1. **Write your name and sign in this page**.
2. Answer each question in the corresponding page.
3. **All answers must be justified**.
4. Personal notes in the formula tables will not be allowed. Over the table it is only permitted to have the DNI (identification document), calculator, pen, and the formula tables.
5. **Do not unstaple any page of the exam** (do not remove the staple).
6. The exam consists of 6 questions, 3 ones corresponding to the first partial (50%) and 3 about the second partial (50%). The lecturer will correct those partial exams indicated by the student with a tick mark in this page. **All questions of each partial exam score the same** (over 10).
7. At the end, it is compulsory to **sign** in the list on the professor's table in order to justify that the exam has been handed in.
8. Time available: **3 hours**

1. (1st Partial) Certain company manufactures two different types of parts (model A and model B). The company is interested in analyzing the time of operation until a failure occurs in the manufacture of these parts. For this purpose, the company has compiled the time until failure (in hours) recorded in recent years. These data are represented in the following plot:



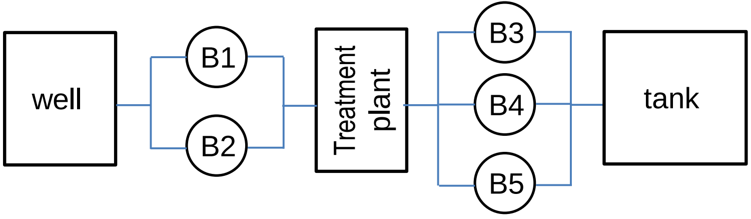
a) What is the population (or populations) under study? What are the individuals? What is the random variable? *(3 points)*

b) The company decides to invest for improving the manufacture of one of the two models. In which model (A or B) would you recommend to invest? Justify your answer. *(2 points)*

c) Indicate if the following statement is true or false, and justify your answer conveniently: "For model B, 50% of failures correspond to a time lower or equal to 72 hours, approximately." *(2 points)*

d) What statistical distribution could be used to reasonably fit the data of model B? Calculate the parameters of this distribution. Based on this distribution, calculate the probability to obtain a time until failure greater than 500 hours in model B. *(3 points)*

2. (1^{st} Partial) In certain industrial plant, two pumps B1 and B2 in parallel conduct water from one well to a treatment plant and, subsequently, other three pumps (B3, B4 and B5, also in parallel) transfer the water to a tank, as indicated in the figure. The lifetime (time of operation until failure) of the treatment plant and the pumps are independent random variables with normal distribution and standard deviation of 6 thousand hours, being 25 thousand hours the mean lifetime of the treatment plant and 20 thousand hours the mean lifetime of each pump.



a) Being T_D the random variable "time of operation until failure of the treatment plant", calculate percentile 7 of this distribution. *(2 points)*

b) Calculate the probability of water arriving from the exit of the treatment plant to the tank after 15 thousand hours of operation. *(2 points)*

c) Calculate the probability of arriving water from the well to the tank after 15 thousand hours of operation. To solve this question, define firstly all events involved in the calculation of the probability. *(3 points)*

d) In order to get a probability exactly of 90% in the previous question (c), what should be the average lifetime of the treatment plant, considering the same standard deviation? *(3 points)*

3. (1st Partial) Certain CRC codes used for sending data packets through the network are capable of correcting a maximum of 7 errors per packet. If this value is exceeded, the packet is rejected. It is known that the average number of errors per packet is equal to 3.

a) What is the probability to reject a packet because it cannot be corrected? *(2 points)*

b) Knowing that a packet has less than 10 errors, what is the probability of being rejected? *(3 points)*

c) If 5 consecutive packets are taken at random, what is the probability that exactly two of them are rejected? *(2 points)*

d) If 5 consecutive packets are taken at random, what is the probability to find a total number of errors less than 18? Solve this question by using the approximation to the normal distribution. *(3 points)*

4. (2nd Partial) It is suspected that certain temperature sensor (S_1) is poorly calibrated. A reference sensor (S_2) properly calibrated by a metrology company is used in order to study this issue. Both sensors are placed inside a climate chamber with unsteady thermo-hygrometric conditions, and one measurement is taken every hour. The values obtained are indicated below:

S_1	18.7	19.4	20.2	21.6	23.8	24.1	26.3	25.7	24.9	24.1	22.9	21.5
S_2	18.4	19.3	20.2	21.8	23.7	24.1	26.5	25.5	24.8	24.1	22.6	21.3
S_1-S_2	0.3	0.1	0	-0.2	0.1	0	-0.2	0.2	0.1	0	0.3	0.2

Average................ 22.7667 (S_1); 22.6917 (S_2); 0.075 (S_1-S_2)
Standard deviation... 2.4810 (S_1) ; 2.5350 (S_2); 0.1658 (S_1-S_2)

a) Based on the values of S_1-S_2, can we accept the null hypothesis that the average value of this variable is null at the population level? Solve this hypothesis test by using the confidence interval method, considering $\alpha = 5\%$. *(3.5 points)*

b) Based on the results obtained in the previous section, is there enough evidence to affirm that sensor S_1 needs to be calibrated? *(1.5 points)*

c) Assuming normality and independence between the values recorded experimentally, obtain a confidence interval for the population variance of sensor S_2 with a confidence level of 95%. *(3 points)*

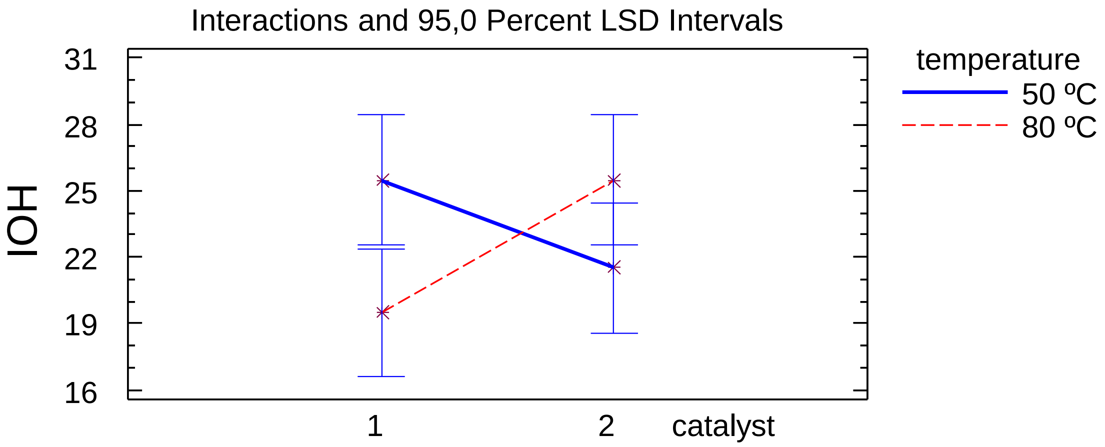
d) If a new measurement is performed inside the climate chamber with sensor S_2, what is the interval that will contain this value with a 95% probability? *Note*: it is assumed that data are independent and they follow a normal distribution, being the mean and standard deviation the values indicated above (below the table) for S_2. *(2 points)*

5. (2º Parcial) A petrochemical company produces certain type of polymer (polyol) by means of a batch process. The hydroxyl index (IOH) is a parameter measured in the final product; this quality parameter is of interest to be as high as possible. An experimental design is carried out to study the possible effect of two factors on the IOH index: temperature inside the reaction tank (two tested levels: 50ºC and 80ºC), and concentration of catalyst (1 g/m^3 or 2 g/m^3). Data obtained experimentally are shown below:

catalyst	2	2	2	2	1	1	1	1
temperature	50	50	80	80	50	50	80	80
IOH	20	23	24	27	24	27	18	21

```
Analysis of Variance for IOH - Type III Sums of Squares
-----------------------------------------------------------------------------
Source                  Sum of Squares     Df     Mean Square    F-Ratio     P-Value
-----------------------------------------------------------------------------
MAIN EFFECTS
 A:catalyst                       2,0      1              2,0       0,44      0,5415
 B:temperature                    2,0      1              2,0       0,44      0,5415

INTERACTIONS
 AB                              50,0      1             50,0      11,11      0,0290
RESIDUAL                         18,0      4              4,5
-----------------------------------------------------------------------------
TOTAL (CORRECTED)                72,0      7
-----------------------------------------------------------------------------
```

Based on the data obtained, an ANOVA was carried out including the double interaction in the model. The interaction plot is shown below:



a) Considering a significance level of 1%, what temperature and what concentration of catalyst would you recommend to achieve a final product with the highest possible IOH value at the population level? *(3 points)*

b) In case of considering α=5%, what temperature and what concentration of catalyst would you recommend? *(3 points)*

c) Describe in detail the procedure used commonly in this type of analysis to detect the presence of outliers. *(2 points)*

d) What useful information could be obtained in this case from the plot of means with LSD intervals for factor "catalyst" in order to study the simple effect of this factor on the IOH parameter at the population level? *(2 points)*

6. (2º Parcial) The resistance of certain polymer used in the production of computer equipment is correlated with the average temperature during the manufacturing process. A set of 100 values of resistance and temperature are available, which have been analyzed by means of simple linear regression with the software *Statgraphics*. The following results were obtained:

```
-----------------------------------------------------------------------------
Multiple Regression Analysis -  Dependent variable: RESISTENCIA
-----------------------------------------------------------------------------
                                            Standard         T
Parameter                   Estimate          Error       Statistic         P-Value
-----------------------------------------------------------------------------
CONSTANT                     13,2566        9,99004        1,32698         0,1876
TEMPERATURA                  1,69615       0,164561        10,3071         0,0000
-----------------------------------------------------------------------------

R-squared = 52,0164 percent
R-squared (adjusted for d.f.) = 51,5268 percent
Standard Error of Est. = 47,3658
```

In order to study if data are better modelled using multiple linear regression, the following table was also obtained:

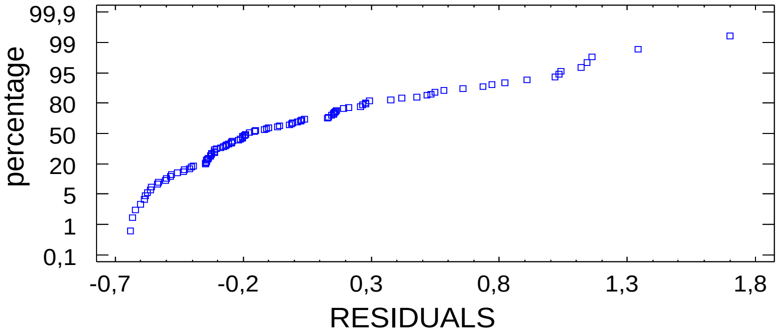
```
-----------------------------------------------------------------------------
Multiple Regression Analysis - Dependent variable: RESISTENCIA
-----------------------------------------------------------------------------
                                         Standard             T
Parameter                   Estimate        Error     Statistic       P-Value
-----------------------------------------------------------------------------
CONSTANT                    -8,61657       16,214     -0,531429        0,5963
TEMPERATURA                  2,85408     0,699248       4,08165        0,0001
TEMPERATURA^2             -0,0108588   0,00637671      -1,70288        0,0918
-----------------------------------------------------------------------------

R-squared = 53,4092 percent
R-squared (adjusted for d.f.) = 52,4486 percent
Standard Error of Est. = 46,9132
```

a) Based on the results, discuss if it is more convenient to use simple linear regression (first model) or multiple regression (second model), considering a significance level of 5%. Justify conveniently your answer. *(2.5 points)*

b) Calculate the correlation coefficient between resistance and temperature from the first model. Will this coefficient be positive or negative? Why? *(2 points)*

c) In the first model, calculate the probability of obtaining a resistance greater than 70 in case of using a temperature of 70ºC. *(3 points)*

d) The residuals of the first model have been represented on a normal probability plot, which is shown below. According to this plot, can we say that there is a non-linear (quadratic) relationship between the variables "*residuals*" and "*percentage*"? What useful information is deduced from this plot? What would you recommend in this case? Justify your answers. *(2.5 points)*



SOLUTION

1a) In this case there are two populations. One is the set of all failures produced in the manufacture of parts from model A: failures produced in the past as well as those that will occur in the future. The other population is the set of failures in the manufacture of parts from model B.
In statistics, a population is the set of all individuals, so that each individual in this case corresponds to a failure in the manufacturing process of these parts.
The random variable is the time (in hours) until a failure occurs in the manufacturing process of the parts, i.e., time since the last failure or time between two consecutive failures.

1b) In a box-whisker plot, the average is represented as a point inside the box, and the median is the line represented inside the box. Based on the plot, it can be deduced that the average as well as the median of the time for model A are lower than for model B. It implies that the manufacturing system of model A fails more frequently than for B, because the time between failures (which is measured by the random variable) are lower in average. Hence, it is recommended to invest in the manufacturing of model A in order to improve the time between failures.

1c) The statement is false because, for model B, 50% of the failures correspond to a time lower or equal to 50 hours, approximately. This value is the median.

1d) As this is a continuous random variable being zero the minimum and the distribution is strongly positively skewed, the values of time between failures could be fitted to an exponential model. This distribution is often used for modeling time until failure or time between failures. This model has only one parameter, which is the inverse of the mean. In this case, the mean is 72 h approximately, so that: $X \approx \exp(\alpha = 1/72)$.
$P(X>500) = e^{-\alpha\cdot 500} = e^{-500/72} = \mathbf{0.0010}$

2a) Considering the units as thousand hours: $T_D \approx N(m=25, \sigma=6)$. By definition, percentile 7 is the value that leaves below 7% of the data. If this value is denoted as k_7:
$P(T_D < k_7) = 0.07$; $P[N(25;6) < k_7] = 0.07$; $P[N(0;1) < (k_7-25)/6] = 0.07$
From the table N(0;1) we obtain that $P[N(0;1) < -1.476] = 0.07$
$(k_7-25)/6 = -1.476 \rightarrow \mathbf{k_7} = 25\text{-}6\cdot 1.476 =$ **16.145 thousand hours**

2b) Let us define the random variable T_{Bi}: time of operation until failure of pump "i". $T_{Bi} \approx N(m=20, \sigma=6) \rightarrow P(T_{Bi} > 15) = P[N(20;6) > 15] =$
$= P[N(0;1) > (15-20)/6] = P[N(0;1) > -0.833] = 1-0.202 = 0.798$

Event B_i: pump "i" is operative more than 15 thousand hours.

P(water from treatment plant to tank) $= P(B_3 \cup B_4 \cup B_5) = 1 - P(\overline{B_3 \cup B_4 \cup B_5}) =$
(by applying one of De Morgan's laws and assuming that events are independent)

$$= 1 - P(\overline{B_3} \cap \overline{B_4} \cap \overline{B_5}) = 1 - P(\overline{B_3}) \cdot P(\overline{B_4}) \cdot P(\overline{B_5}) = 1 - (1-0.798)^3 = \mathbf{0.9917}$$

2c) Event B_i: pump "i" is operative more than 15 thousand hours.
Event $B_{1\text{-}2}$: pump B1 or B2 are operative more than 15 thousand hours.
Event $B_{3\text{-}4\text{-}5}$: at least one of the 3 pumps are operative more than 15,000 hours.
Event D: the treatment plant is operative more than 15 thousand hours.
Event w→t: the water is carried from the well to the tank after 15,000 hours.

$P(B_{3\text{-}4\text{-}5}) = 0.9917$ (computed in the previous section)
$P\ (B_{1\text{-}2})$ = calculation analogous to previous section, replacing the exponent 3 by 2 =

$$= P(B_1 \cup B_2) = 1 - (1 - 0.798)^2 = 0.9591.$$

$P(D) = P[N(25;\ 6) > 15] = P[N(0;\ 1) > (15\text{-}25)/6] = P[N(0;1) > \text{-}1.667] = 0.952$

Assuming that events are independent: $P(w \to t) = P(B_{1-2} \cap D \cap B_{3-4-5}) =$
$= P(B_{1\text{-}2}) \cdot P(D) \cdot P(B_{3\text{-}4\text{-}5}) = 0.9591 \cdot 0.952 \cdot 0.9917 = \mathbf{0.9057}$

2d) $P(B_{1\text{-}2}) \cdot P(D) \cdot P(B_{3\text{-}4\text{-}5}) = 0.9591 \cdot P(D) \cdot 0.9917 = 0.90 \quad \to P(D) = 0.9462.$
$P(D) = P[N(m;\ 6) > 15] = 0.9462;\ \ P[N(0;\ 1) > (15\text{-}m)/6] = 0.9462;$
From the table N(0; 1) it is obtained that:
$(15\text{-}m)/6 = \text{-}1.609 \to m = 15 + 1.609 \cdot 6 \to m =$ **24.65 thousand hours**
As the probability computed in the previous questions is nearly 90%, the average lifetime of the treatment plant is nearly the initial value, i.e., close to 25 thousand hours.

3a) Let us define X: number of errors of one packet. In this discrete random variable, the minimum is zero and the maximum is not determined; hence, it can be modelled by means of a Poisson distribution, whose parameter is the mean: $X \approx Ps\ (\lambda=3)$. P(reject packet) = $P(X>7) = 1\text{-}P(X \leq 7) = 1\text{-}\ 0.988 = \mathbf{0.012}$
This value is obtained from the abacus, placing a vertical line for $\lambda=3$ (horizontal axis) and crossing the curve $\nu=7$.

3b) We have to compute a conditional probability, by applying the formula:

$$P[(X > 7)/(X < 10)] = \frac{P[(X>7)\cap(X<10)]}{P(X<10)} = \frac{P(X=8)+P(X=9)}{P(X<10)} = \frac{0.008101+0.0027}{0.9989} = \mathbf{0.0108}$$

$P(X<10) = P(X \leq 9) = 0.9989$ (value from the abacus, using $\lambda=3$).
By applying the probability function for the Poisson distribution, it turns out:
$P(X=8) = e^{-3} \cdot 3^8/8! = 0.0081;\quad P(X=9) = e^{-3} \cdot 3^9/9! = 0.0027$

3c) Let us define Y: number of packets rejected out of a set of 5 packets. This discrete random variable ranges from zero to 5 and, hence, it can be modelled as a Binomial distribution with parameters: n=5 and p = probability to reject one packet = 0.012 (from section *a*) → $Y \approx Bi\ (n=5,\ p=0.012)$
In order to calculate the probability that this variable takes the value 2, we have to apply the probability function:

$$P(Y=2) = P[Bi(5;0.012) = 2] = \binom{5}{2} \cdot 0.012^2 \cdot (1 - 0.012)^{(5-2)} = \mathbf{0.00139}$$

The combinatorial number is: $5!\ /\ (2! \cdot 3!) = 120/2 \cdot 6 = 10$

3d) Let us define Z: total number of errors contained in 5 packets: Z = $X_1 + X_2 + \ldots + X_5$. The sum of independent Poisson random variables follows a Poisson distribution, being the parameter the sum of λ_i: Z ≈ Ps (λ=5·3 = 15). As this parameter is greater than 9, the probability function resembles a Gaussian model (normal distribution), with mean =15 and variance =15. The correction of continuity has to be applied because we are approximating a discrete distribution by means of a continuous distribution:

$$P(Z<18) = P[Ps\ (15) < 18] \approx P\left[N\left(15, \sqrt{15}\right) < 17.5\right] = P\left[N(0;1) < \frac{17.5-15}{\sqrt{15}}\right] =$$

$$= P[N(0;\ 1) < 0.6455] = 1 - 0.259 = \mathbf{0.741}$$

4a) The hypothesis test is H_0: $m_{S1\text{-}S2} = 0$; H_1: $m_{S1\text{-}S2} \neq 0$.
Sample mean of S_1-S_2 = 0.075; std. deviation = 0.1658; n=12. ; α=0.05
The critical value of a Student's t distribution with 11 degrees of freedom that leaves above an area of 0.025 is 2.201.

$$m \in \left[\bar{x} - t_{n-1}^{\frac{\alpha}{2}} \cdot s/\sqrt{n}\,;\ \bar{x} + t_{n-1}^{\frac{\alpha}{2}} \cdot s/\sqrt{n}\right]\ ;\quad m \in \left[0.075 \pm 2.201 \cdot 0.1658/\sqrt{12}\right]\ ;$$

$m \in [\mathbf{-0.0304; 0.1804}]$ Given that the value zero is contained in this interval, we can admit the null hypothesis that the population mean of S_1-S_2 is zero.

4b) As it can be accepted that $m_{S1\text{-}S2}$ = 0, it can be deduced that: m_{S1} = m_{S2}. Thus, there is not enough evidence to affirm that the average at the population level of measurements from sensor S_1 is different to the population average of measurements from S_2. There is not enough evidence to say that sensor S_1 requires to be calibrated, because such calibration is necessary when measurements from one sensor, on average, differ significantly from those obtained with a calibrated sensor.

4c) Confidence interval for the population variance of S_2, with 1-α=95%:

$$\sigma^2 \in [(n-1) \cdot s^2/g_2\,; (n-1) \cdot s^2/g_1]\ \ ;\sigma^2 \in [11 \cdot 2.535^2/21.92\,; 11 \cdot 2.535^2/3.816]$$

$$\boldsymbol{\sigma^2} \in [\mathbf{3.225\ ; 18.525}]$$

Being g_1=3.816 y g_2=21.920 the interval of a chi-squared distribution with 11 degrees of freedom that comprises 95% of the values. The sample standard deviation is s = 2.535 (value indicated in the statement).

4d) In a normal distribution, the interval $m \pm 1.96 \cdot \sigma$ comprises 95% of the values. Considering m=22.3917 and σ=2.535, it turns out:

$22.6917 \pm 1.96 \cdot 2.535$ → **[17.723; 27.660]**
This interval will comprise 95% of all measurements carried out inside the climatic chamber with sensor S_2.

5a) Using α=1%, the interaction is not statistically significant because the p-avlue associated to it (0.029) is greater than 0.01. Also, the simple effect of both factors is not statistically significant, for the same reason. Thus, there is not enough evidence to affirm that it is better to use a temperature of 50ºC or 80ºC, and it is also uncertain which concentration of catalyst, 1 or 2 g/m^3, would lead to a higher value of IOH on average at the population level. The recommendation would be to use the cheapest operative conditions: a concentration of 1 g/m^3, because obviously it becomes cheaper than using a double amount, and a temperature of 50ºC which will be cheaper than having to reach 80ºC inside the reaction tank.

5b) Considering α=5%, the interaction becomes statistically significant (p-value = 0.029 < 0.05). Thus, the effect of changing the temperature is significantly different according to the concentration of catalyst. For conc=1, LSD intervals (obtained using 1-α=95%) do not overlap for both temperatures; hence, it is of interest to use temperature=50ºC in order to maximize the response variable IOH at the population level. However, LSD interval for conc=1 and temp=50 overlaps with the two intervals for conc=2, which implies that there is not enough evidence to affirm that conc=1 and temp=50 is more convenient that conc=2 at the population level. Nonetheless, given that conc=1 is the cheapest condition for obvious reasons, that would be the recommended option (conc=1 and temp=50ºC).

5c) The procedure to detect outliers in this type of data analysis (ANOVA with two factors) is the following:
1) If the double interaction is statistically significant, it must be included in the model. If it is not significant, it should be discarded, as well as the factors whose simple effect is not significant.
2) Subsequently, it is necessary to obtain the residuals of the ANOVA model, and to plot them on a normal probability plot.
3) If the points fit reasonably well to a straight line but an extreme value appears clearly separated from the line, it will correspond to an outlier.
4) If the points do not fit a straight line and reveal an asymmetric distribution of residuals, the original data should be transformed and, next, this procedure should be repeated (i.e., go to step "1").

5d) Any useful information: as the simple effect of factor "catalyst" is not statistically significant (p-value = 0.54 > α), it is known that the plot of means with LSD intervals will show the two intervals with certain overlap. Thus, this plot does not provide any useful information in this case for inference purposes.

6a) The multiple regression model is: $Y = \beta_0 + \beta_1 \cdot \text{Temp} + \beta_2 \cdot \text{Temp}^2$. This equation will be more appropriate than the simple linear regression if it can be guaranteed that coefficient β_2 is different from zero at the population level. We have to test the null hypothesis H_0: $\beta_2 = 0$ against the alternative H_1: $\beta_2 \neq 0$. The p-value associated to this hypothesis test is 0.0918, which is greater than 0.05 (significance level) and, hence, we can accept the null hypothesis. In the quadratic equation, if coefficient β_2 is considered as zero, it becomes model 1 (simple regression). As there is not enough evidence to affirm a quadratic effect at the population level, it is of interest to use the first model.

6b) In simple regression, the correlation coefficient is computed as the square root of the determination coefficient: $r = \sqrt{R^2} = \sqrt{0.52016} = \mathbf{0.7212}$

This coefficient is positive in this case because the slope of the line is positive (in case of a negative slope, we should put a "minus" sign). A positive correlation implies that when the temperature in the manufacturing process is higher, the polymer resistance also becomes higher on average.

6c) Assuming the hypothesis of normality and homoscedasticity, the conditional distribution of resistance (Y) when temperature is 70ºC will be a normal model with mean estimated using the regression equation and with standard deviation of 47.3658 (value indicated in the table as *standard error of est.*): Equation of the model: **Y = 13.2566 + 1.69615 · Temp**. The constant cannot be removed although it is not statistically significant, because the model should be recalculated and the new value estimated for the slope is unknown.

E(Y/X=70) = 13.2566 + 1.69615·70 = 131.99

P[N(131.99, 47.366) >70] = P[N(0; 1) > (70-131.99)/47.366] =
= P[N(0;1) > -1.309] = 1-0.0951 = **0.9047**

This calculation is approximate because the graph shown in the next question reveals a positively skewed distribution of the residuals; hence, the hypothesis of normality is not accomplished.

6d) 1. It is nonsense to affirm a quadratic relationship between "residuals" and "percentage" because this normal probability plot represents the data of a single one-dimensional random variable, it does not intend to visualize the relationship between two variables (as it would be the case of a scatterplot). The vertical axis represents the cumulative frequency (in percentage) of residuals in a special scale.

2. Useful information: the points in the plot do not fit to a straight line, they follow a curvature. This indicates a positively skewed distribution of residuals. There is not enough evidence to affirm the presence of outliers that should be discarded.

3. In this case it is recommended to transform the data (resistance values), to repeat the regression model, re-calculate the residuals, repeat the normal probability plot again and check if the new residuals tend to be normal and if any outlier is identified. The most common transformations are the square root ($x^{1/2}$), logarithm or fourth root ($x^{1/4}$).