Bachelor Degree in Computer Engineering

Statistics

FINAL EXAM

June 20th 2017

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) One study has been carried out about the telecommunication services in 770 villages of a Spanish region in certain year. For each village, the following characteristics were determined:

- Number of ADSL lines under operation.
- Average age (years) of the population living in the village.
- Total population of the village.
- Average family income per person (€), coded as indicated below:

Frequency Table for family income (average for the village) (Table 1)

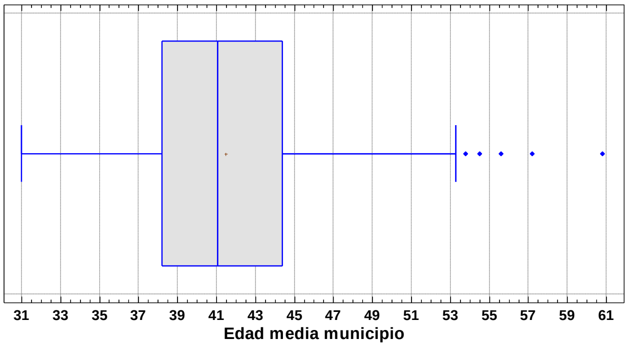
Class	*Value*	*Frequency*	*Relative frequency*	*Cumulative frequency*	*Cum. Rel. frequency*
1	Below 7,200	141	0.1831	141	
2	From 7,200 to 8,300	308	0.4000	449	
3	From 8,300 to 9,300	202	0.2623	651	
4	From 9,300 to 10,200	93	0.1208	744	
5	From 10,200 to 11,300	22	0.0286	766	
6	From 11,300 to 12,100				
7	From 12,100 to 12,700	1	0.0013		

A descriptive study of the data has led to the following results:

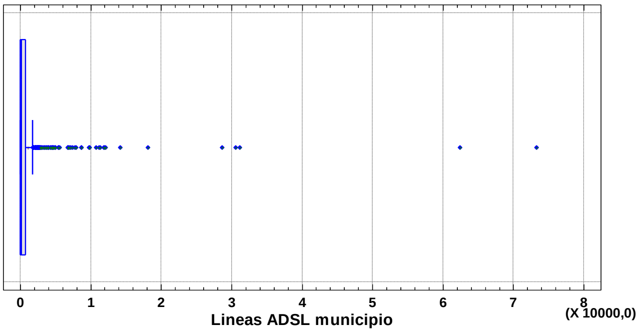
Summary Statistics (Table 2)

	Average age	*ADSL lines*	*Population*	Type
Count	770	770	770	
Average	41.4627	1075.84	10466.8	
Median		181.0	2761.0	
Standard deviation	4.4481	4276.64	39302.0	
Coef. of variation		397.517%	375.49%	
Minimum	31.0		50.0	
Maximum		73274.0	699145	
Range		73274.0	699095	
Lower quartile	38.2	26.0	1012.0	
Upper quartile	44.4	736.0	7054.0	
Interquartile range	6.2		6042.0	
Standard skewness	1.9778		136.523	
Standard kurtosis	1.24895	953.325	1018.47	

Plot 1



Plot 2



According to the results shown above (parameters, tables and plots), answer the following questions justifying the reply:

a) Fill in the empty cells (in dark) in Table 2 (summary statistics) with the appropriate values, and indicate inside the table (column on the right) the type of each parameter. *(2,5 points)*

b) What is the type of the variables studied? What is their dimension? *(2 points)*

c) Indicate the most representative parameters of position and dispersion for the variables "average age", "ADSL lines" and "population". Justify your answer. Which one has the highest dispersion? *(1,5 points)*

d) Regarding the family income, calculate the percentage of villages in the region under study with a family income below or equal to 12,100 €. *(1 point)*

e) Regarding the family income, calculate the number of villages in the region with a family income between 11,300 and 12,100. *(1 point)*

f) What is the name of plots 1 and 2? Explain the advantages and disadvantages of this type of plot with respect to histograms. *(1 point)*

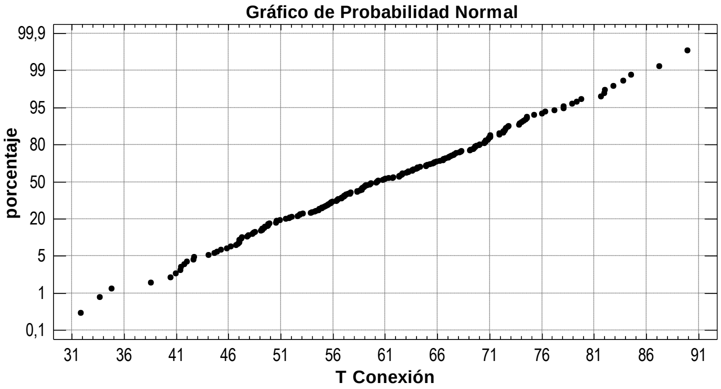
g) What type of graphical representation might be the most suitable to describe the distribution of "family income"? Draw approximately this representation and mark graphically the basic parameters. *(1 point)*

2. (1^{st} Partial) One hotel in Quiet Bay has 120 rooms. During spring months, the daily hotel occupation is about 75%. If the number of pre-booked rooms exceeds 85% of the total number of rooms at the hotel, the management has to hire extra staff to attend the customers. It is assumed that all spring days are equally "desirable" by customers to make a reservation.

a) What is the probability to have all rooms occupied a given day of spring? *(2 points)*

b) What is approximately the probability to have to hire extra staff a given day of spring? *(4 points)*

c) The hotel is planning to purchase a new router to improve the WiFi connection service. In order to determine the characteristics of the new router, the hotel has recorded the time of WiFi connection for a sample of customers. These values were represented on the Normal Probability Plot shown below. According to the plot, answer the following questions:



c.1) Calculate, approximately, the average connection time. *(1.5 points)*

c.2) Calculate the percentage of connections with a time between 50 and 70 minutes. *(2.5 points)*

3. (1st Partial) The time (T) of operation until failure (in hours) of certain electronic components fluctuates randomly according to an exponential distribution. Experimental evidence indicates that 10% of components fail before one year.

a) Calculate the average time of operation (half-life) of this type of components. *(3 points)*

b) If one component has been operative during 3 years without suffering any failure, what is the probability to be operative during 5 additional years? *(3 points)*

c) These components are used in the assembly of certain electronic device (D) that is connected to two terminals A and B. The target is to reach a reliability at least of 90% at one year. In order to ensure this reliability, N components of this type are assembled in parallel. Calculate the minimum value N, assuming independence in their operation or failure. *(4 points)*

4. (2nd Partial) Answer these questions justifying conveniently the reply.

a) One sample of size 20 is randomly taken from a normal population with standard deviation σ = 4. What is the probability to obtain a sample mean that differs less than one unit with respect to the population mean? *(3 points)*

b) One quality parameter is routinely measured in a process. Certain operational changes are carries out to improve the quality. In order to study the effectiveness of such changes, a sample of 18 units is randomly taken, resulting a sample mean of 21 and a sample variance of 2.3. Obtain a confidence level for the population mean, considering α=0.01. Is it reasonable to assume a value of 22 for the population mean of this parameter, after the operational changes carried out? *(3.5 points)*

c) Is it acceptable to assume a value of 8.0 units^2 for the population variance of this parameter, after the operational changes carried out? (α=0.01) *(3.5 points)*

5. (2nd Partial) One experiment was carried out to study the effect of two factors on the average performance of a computer system. The factors are: *configuration* (three possibilities: A, B or C) and *size of Cache memory* (3 levels: low, medium or high). Each treatment was tested twice.

a) Obtain the summary table of ANOVA (use α=5%) taking into account that: SS_{total}=11039.2; SS_{config}=704.47; SC_{cache}=8390.4; $SC_{residual}$=690.005. What effects are statistically significant? Indicate all the hypotheses tested with ANOVA with respect to the effects studied. *(5 points)*

b) How many populations are being compared in this study? Assuming that the hypothesis of homoscedasticity is fulfilled, calculate the estimated variance for each one of these populations. *(2 points)*

c) Generally speaking, what additional information, with respect to that provided by the summary table of ANOVA, can be obtained from the graphical representation of LSD intervals? *(3 points)*

6. (2nd Partial) Answer the following questions A and B:

A).- The mean daily load of a computer system (X, measured as number of queries per minute) has been recorded during 3 months in order to study the performance of certain database, as well as the average response time in seconds (Y). The following parameters were obtained from the data gathered in the study: $\bar{X}$= 3.5 S_X= 0.65 $\bar{Y}$=1.3 S_Y=0.6 r_{XY} = 0.9.

a) Based on this information, write the equation of the regression line that allows to estimate the average response time according to the number of queries. What is the practical interpretation of the numeric value of the slope? *(3 points)*

b) Calculate the coefficient of determination. What is the practical interpretation of the value obtained? *(1 point)*

c) If the load of certain day is 7 queries per minute, calculate the probability to obtain a response time greater than 5 seconds. *(3 points)*

B).- A correlation coefficient -0.82 was found between the number of people suffering influenza, reported weekly in a city, and the amount of beer sold, for a period of 5 years.

B.1) This study seems to suggest that the consumption of beer helps to prevent influenza. What do you think about this statement? *(1,5 points)*

B.2) The study has also calculated the correlation between the amount of beer consumed during the first week of August by 25 families from that city and 25 families from another neighboring city. In this case, the correlation coefficient turned out to be positive. What can be concluded about this result? *(1.5 points)*

SOLUTION

1a) Parameters about Age: Median = **41.1** (vertical line inside the box).
Coefficient of variation = std. dev. / mean = 4.448/41.463 = **0.107**
Maximum = **60.9** (appears as an isolated point in the plot).
Range = maximum - minimum = 60.9 - 31 = **29.9**

Parameters about ADSL lines: minimum = max - range = 73,274 - 73,274 = **0**
Interquartile range = Q_3 - Q_1 = 736 - 26 = **710**
Standard skewness: >> **2** (the exact value is unknown, but plot 2 reveals that the distribution is strongly positively skewed).

Type of each parameter:

- Position: average, median, minimum, maximum, lower quartile, and upper quartile.
- Dispersion: std. deviation, coeff. of variation, range, interquartile range.
- Shape: standard skewness and standard kurtosis.

The parameter "count" cannot be classified into any of these categories.

1b) Four characteristics are known for each individual of the population (village): number of ADSL lines, average age, total population, and family income. Thus, we are dealing with a **four**-dimensional random variable:

- ADSL lines and total population: both are quantitative random variables containing discrete values.
- Average age: quantitative random variable containing continuous values, because it is an average of discrete values.
- Average family income per person: it is a continuous variable, but in this case the values are coded into seven categories. Thus, it can be considered as a quantitative random variable coded as a discrete set of categories.

1c) The distribution of "average age" is quite symmetric (plot 1) and it can be considered as a random sample from a normal distribution because the standard skewness is comprised between -2 and 2. In a normal distribution, the most representative parameter of position is the mean (which is coincident with the median), and for the dispersion, the most representative parameters are the variance or standard deviation.

The distribution of ADSL lines is strongly positively skewed according to plot 2, as well as total population (because the standard skewness is 136, much greater than 2). In such cases, the median is more representative than the mean as parameter of position because it is not affected by extreme values. For the same reason, the interquartile range is the most representative parameter of dispersion.

Population is the variable with highest dispersion because its interquartile range (6042) is much greater than for ADSL (710) as well as "age" (6.2). The same applies with the standard deviation.

1d) Percentage of villages with a family income ≤ 12,100 € =
= 100·(770-1)/770 = 100·(1-0.0013) = **99.87%**

1e) Number of villages with a family income between 11,300 and 12,100 =
= 770 - (141+308+202+93+22+1) = 770 - 767 = **3**.

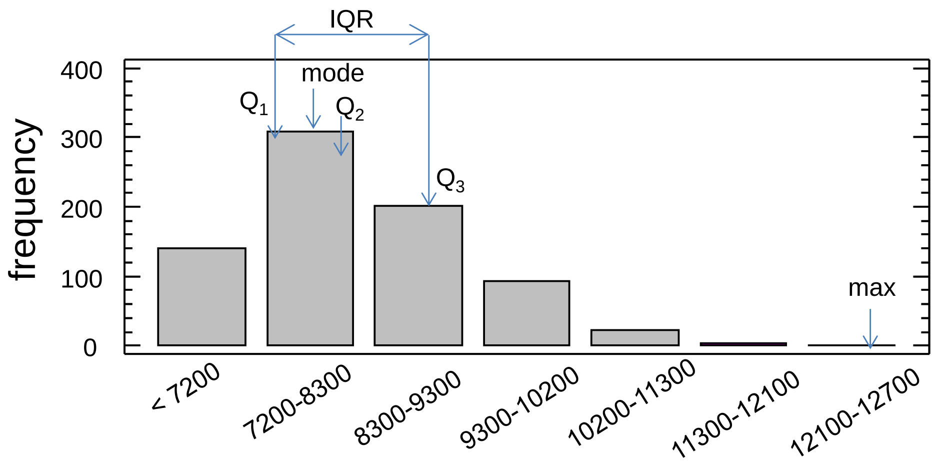
1f) Plots 1 and 2 are called box-and-whisker plots. Advantages:

- They are useful to visualize extreme values, which is helpful to diagnose outliers or abnormal data.
- In case of few values, they are more convenient than histograms in order to discuss the symmetry or skewness of the distribution.
- They visualize key parameters like mean, median and quartiles, which allows the direct calculation of interquartile range.
- Their shape is constant, it does not depend on the number of classes chosen, like in the case of histograms.
- The comparison of several box-whisker plots is easy (multiple box-whisker plots), which does not apply for histograms.

Disadvantages of box-whisker plots with respect to histograms:

- The vertical scale does not provide information. Thus, it is not possible to know the total number of values, or the number of data within each interval: frequencies or percentiles cannot be calculated.
- Histograms are able to reveal certain patterns of variability more clearly than box-whisker plots, for example bimodal distributions, truncated data or abnormal frequencies of certain values.

1g) Data of family income are coded into 7 categories. Thus, a **bar-chart** would be the most suitable graphical representation to describe the distribution of this random variable, which is depicted below. The approximate position of the basic parameters of the distribution (first quartile, mode, median, third quartile, interquartile range (IQR) and maximum are marked. In a frequency histogram, all bars have the same width, which is not the case here. Thus, it is not possible to draw a histogram in this case.



2a) The daily hotel occupation of 75% is *not* the probability that the hotel is fully booked in a given day. It means the probability of a room in the hotel to be booked: if one day of spring is chosen, 75% of the rooms are booked on average. Thus, the random variable X is defined as "number of rooms booked in a hotel with 120 rooms". This variable ranges from 0 to 120 and, hence, it follows a binomial distribution: B (n=120, p = 0.75). The probability to have all rooms occupied is obtained by applying the probability function:

$$P(X=120)=\binom{120}{120}\cdot 0.75^{120}\cdot(1-0.75)^{0}=0.75^{120}=\mathbf{1.02\cdot 10^{-15}}$$

2b) X≈B (n=120; p=0.75); E(X) = n·p = 120·0.75 = 90

$$\sigma_x=\sqrt{n\cdot p\cdot(1-p)}=\sqrt{120\cdot 0.75\cdot 0.25}=\sqrt{22.5}=4.743$$

As the variance is greater than 9, the distribution can be approximated by means of a Normal: X ≈ N (m=90, σ=4.743). If the number of pre-booked rooms exceeds 85% of the total (i.e., 0.85·120 = 102), the management has to hire extra staff. Thus, probability to hire extra staff = P(X>102):

$$P(X>102)\approx P[N(90;\ 4.74)>102.5]=P\left[N(0;\ 1)>\frac{102.5-90}{4.743}\right]=P[N(0;1)>2.635]=0.0042$$

2c.1) As the points follow approximately a straight line, it can be concluded that data are normally distributed. The vertical scale gives the probability: P(X≤x). According to the plot, P(X≤60) = 50%, which implies that 60 is the median. But as the distribution is normal, the average will be around **60**.

2c.2) By reading 80% on the vertical scale, the horizontal line crosses the line of points approximately at X = 70, which means that P(X≤70)=80%. Similarly, it turns out from the plot that P(X≤50) ≈ 15%. Based on this information:

$$P(X\in[50,\ 70])=P(X\le 70)-P(X\le 50)\approx 80\%-15\%\approx\mathbf{65\%}$$

3a) The random variable is defined as T: time (in hours) of operation until failure. Let's define another variable X: time (in years) of operation until failure.

X ≈exp (α); $P(X>1)=0.9=e^{-\alpha\cdot 1}$; ln 0.9 = -α ; α = 0.1054; E(X)=1/α = 9.49 years

E(T) = 9.49 years x 365 days/year x 24 hours/day = **83,143 hours**

3b) The exponential distribution satisfies the lack-of-memory property:

P(X>8 / X>3) = P(X>5) = $e^{-0.1054\cdot 5}$ = **0.59**

Another way to solve the problem is by applying the conditional probability:

$$P[(X>8)/(X>3)]=\frac{P[(X>8)\cap(X>3)]}{P(X>3)}=\frac{P(X>8)}{P(X>3)}=\frac{e^{-0.105\cdot 8}}{e^{-0.105\cdot 3}}=0.59$$

3c) Random variable Y: time of operation of the device until failure. Target of the problem: reliability of the device ≥ 0.9 at one year. According to the concept of reliability, the requirement is: P(Y≥1) ≥ 0.9. But for each individual component, 10% of them fail before one year: P(X<1) = 0.1 → P(X≥1)= 0.9. Thus, just with one component the requirement is already satisfied: the solution is **N=1**.

4a) If a sample of size n=20 is randomly taken from a normal population with σ=4, the distribution of the sample mean is:

$$\bar{x} \approx N\left(m;\sigma/\sqrt{n}\right) \approx N\left(m;\ \ 4/\sqrt{20}\right) \quad \to \quad \bar{x}-m \approx N\left(0;\ \ 4/\sqrt{20}\right)$$

What is the probability to obtain a sample mean that differs less than one unit with respect to the population mean? The difference is in absolute value:

$$P\left(\left|\bar{x}-m\right|<1\right)=P\left[\left(\bar{x}-m\right)\in[-1;\ 1]\right]=1-2\cdot P\left[\left(\bar{x}-m\right)>1\right]=1-2\cdot P\left[N\left(0;\ \ 4/\sqrt{20}\right)>1\right]=$$
$$=1-2\cdot P\left[N\left(0;\ 1\right)>\sqrt{20}/4\right]=1-2\cdot P\left[N\left(0;\ 1\right)>1.118\right]=1-2\cdot 0{,}131=0.736$$

4b) In this case, n=18, s^2=2.3, $\bar{x}=21$; $t_{17}^{0.005}=2.898$. Confidence interval for m:

$$CI_m=\bar{x}\pm t_{n-1}^{\alpha/2}\cdot\frac{s}{\sqrt{n}}=21\pm 2.898\cdot\frac{\sqrt{2.3}}{\sqrt{18}}=21\pm 1.036=\left[19.96;\ 22.04\right]$$

As the value 22 is inside the interval obtained, it is reasonable to assume a value of 22 for the population mean (the null hypothesis m=22 is acceptable).

4c) Given α=0.01, we obtain from the chi^2 table that: $P(\chi_{17}^2>35.718)=0.005$. Moreover, $P(\chi_{17}^2<5.697)=0.005$. These are the critical values used to obtain the confidence interval (CI) for the population variance:

$$CI_{\sigma^2}=\left[\frac{(n-1)\cdot s^2}{g_2};\frac{(n-1)\cdot s^2}{g_1}\right]=\left[\frac{17\cdot 2.3}{35.718};\frac{17\cdot 2.3}{5.697}\right]=\left[1.095;\ 6.863\right]$$

As the presumed value for H_0: σ^2 = 8 is outside the interval obtained, it is **not acceptable** to assume a value of 8 for the population variance considering α=0.01.

5a) There are 9 treatment (3 types of configurations combined by 3 types of cache memory). As each treatment was tested twice, the total number of values (N) is 18. Total degrees of freedom (d.f.) = N-1 = 17; D.f. for each factor = number of variants -1; D.f. interaction = 2·2 = 4; $SS_{interac}$ = SS_{total} - SS_{config} - SS_{mem} - SS_{resid}. Mean square = SS / d.f. F-ratio = MS / MS_{resid}.

	SS	d.f.	MS	F-ratio
Configuration	**704.47**	2	352.235	4.59 > ($F_{2;9}^{0.05}=4.26$)
Memory	**8,390.40**	2	4,195.2	54.72 > ($F_{2;9}^{0.05}=4.26$)
Interaction	1,254.325	4	313.58	4.09 > ($F_{4;9}^{0.05}=3.63$)
Residual	**690.005**	9	46.667	
Total	**11,039.20**	17		

- For configuration: H_0: $m_A = m_B = m_C$. As the F-ratio is greater than the critical value obtained from the F table for α=0.05, the null hypothesis is rejected, which means that the mean value of at least one of the configurations is different from the rest at the population level. Thus, it can be concluded that the simple effect of configuration is statistically significant.

- For cache memory: H_0: $m_{low} = m_{medium} = m_{high}$. As the F-ratio is greater than the critical value, H_0 is rejected: the mean value of at least one of the memories is different from the rest at the population level. Thus, the simple effect of cache memory is statistically significant.

- For the double interaction: the null hypothesis is that the effect of the double interaction is zero at the population level (i.e., the effect on the response variable of changing the configuration does not depend on the cache memory). As the F-ratio is greater than the critical value, H_0 is rejected: the effect of the double interaction is statistically significant.

5b) The number of populations is equal to the number of treatments = **9**, because the average performance of each treatment can be characterized by a different average. As the variance of all populations is assumed to be the same (hypothesis of homoscedasticity), this variance can be estimated by means of the mean squares of the residuals obtained in the ANOVA table: MS_{resid} = **76.67**.

5c) If the ANOVA table reveals that the simple effect of a factor is not statistically significant, then LSD intervals do not provide additional information (the plot will show that all intervals are overlapped). In the case of qualitative factors with more than two variants, if the effect of a factor is statistically significant according to the ANOVA table, it can be concluded that at least two of the means will be different at the population level, but LSD intervals are necessary in order to determine which variants are the ones with a different mean, which will be those whose LSD intervals will not be overlapped.
In the case of quantitative factors with more than two levels, the plot of LSD intervals provides clues to interpret the nature (linear or quadratic) of the effect, though an additional analysis using linear regression will be necessary to study the statistical significance of the relationship.

6a) Slope: $b = r \cdot s_y / s_x = 0.9 \cdot 0.6 / 0.65 = 0.8306$

Intercept: $a = \overline{y} - b \cdot \overline{x} = 1.3 - 0.8306 \cdot 3.5 = -1.608$

Regression model: **Time = -1.608 + 0.8306 · $N_{queries}$**
As the correlation coefficient is rather high, this equation will be useful to predict the average time of response as a function of the number of queries per minute.

Interpretation of the slope: the value 0.83 indicates that the dependent variable (response time) will increase in 0.83 seconds (in average) when the number of queries increases one unit.

6b) Coefficient of determination: $R^2 = 100 \cdot r^2 = 100 \cdot 0.9^2 =$ **81%**
This coefficient represents the percentage of variation of the Y variable explained by the variability of X. Thus, it turns out that 81% of the variability of the response time is explained by the model, i.e., by the variable “number of queries”.

6c) The conditional distribution of time when x=7 is a normal distribution with the following parameters:

$$E(Y/X=7) = -1.608 + 0.8306 \cdot 7 = 4.2062$$

$$\sigma(Y/X=7) = \sigma_{resid} = \sqrt{s_Y^2 \cdot (1-r^2)} = \sqrt{0.6^2 \cdot (1-0.9^2)} = \sqrt{0.0684} = 0.2615$$

$$P[(Y>5)/(X=7)] = P[N(4.206;\ 0.261)>5] = P\left[N(0;1) > \frac{5-4.206}{0.2615}\right] = P[N(0;1)>3.036] = 0.0012$$

6B.1) The rather strong correlation coefficient indicates that a higher amount of beer sold in the city corresponds to those weeks with lower people suffering influenza. This observed correlation might suggest that a higher consumption of beer prevents people from suffering influenza. However, this conclusion might not be true because this is not a cause-effect relationship. This is an example of spurious correlation: *a mathematical relationship in which two or more variables are not causally related to each other, yet it may be wrongly inferred that they are, due to either coincidence or the presence of a certain third, unseen factor (referred to as a “common response variable” or “confounding factor”).* Further information at: https://en.wikipedia.org/wiki/Spurious_relationship

In this case, the confounding factor might be the different temperature along the year: during the hot months, the reported cases of influenza are obviously lower and, due to the warm weather, people like meeting at bars and drinking beer. Conversely, during the cold season, the cases of influenza are much higher and people tend to drink less beer.

6B.2) There are 25 families from one city and 25 different families from another city. There is not any relationship between the families, which implies that data are structured as two one-dimensional variables. In such case, a simple regression of both variables is nonsense. The positive correlation is obtained at random, and for sure the value will not be very high.