EXAM 1: PROBABILITY

14th April 2017

Question 1. In Spain, one of the most popular lotteries is called “La Primitiva”. Participants have to choose 6 different numbers between 1 and 49. Additionally, a random number between 0 and 9 (that goes apart) and that is called “reintegro” is assigned to each participant. Its only function is that those people guessing the “reintegro” will receive back the money they have paid.
The drawing goes like this: six different numbers are chosen at first between 1 and 49; then, a seventh different number between 1 and 49 is also selected, and this is called the “complementario”. Finally, a number between 0 and 9 is chosen for the “reintegro”. We will forget about it in this problem.

- Find the probability of winning the big prize of “La Primitiva” (that happens when your six numbers match with the six numbers that have been selected in the drawing, without taken into account the “complementario”).
- Find the probability of winning the second prize, that happens when you have 5 out of 6 correct numbers and the other number you have agrees with the “complementario”.
- Find the probability of winning the third prize, that happens when you have 5 out of 6 correct numbers and the other does not agree with the “complementario”.
- How many times you have to participate in “La Primitiva” if you want to have more than 50% of chances of winning the big prize?

Question 2. We have three urns with 5 balls each, with the following composition (W for white and B for black):

$$U_1(3W, 2B), \quad U_2(2W, 3B), \quad U_3(1W, 4B).$$

Two dices are rolled. If the sum obtained is smaller or equal than 6, we choose the first urn. If the sum is 7 we choose U_2; otherwise we choose U_3. Then, a ball is picked up randomly from the chosen urn.

- Find the probability of choosing the first urn.
- Find the probability that the chosen ball is white.
- Knowing that the chosen ball is white, determine the probability that the sum of points obtained in the dices were 7.

Question 3. A certain random variable X takes 3 different values, $6, 10, 12$, with the following probability function

$$\Pr(X = 6) = t^2, \quad \Pr(X = 10) = t, \quad \Pr(X = 12) = t/2,$$

where t is a certain real number.

- Find the values of t such that what we have defined is really a probability function. For instance, $t = 0$ is not possible since this would imply that the sum of probabilities is not one.
- For the t you have found, determine the expectation of X.
- For the t you have found, determine the variance of X.

Question 1.

- There are $\binom{49}{6}$ of picking up 6 distinct numbers in a set of 49; just one of these combinations is good, so the probability is just $\frac{1}{\binom{49}{6}}$.
- For the second prize, five out of six numbers must agree with those of the drawing. This is done in $\binom{6}{5} = 6$ ways; the other must agree with the "complementario", so there are 6 possibilities for having the second prize, and the probability is $\frac{6}{\binom{49}{6}}$. Alternatively, the probability of getting 5 out of 6 correct numbers is $\frac{6\cdot 43}{\binom{49}{6}}$, that you must multiply by the probability that later your other number agrees with the "complementario", that is equal to $1/43$, since at the moment of deciding that number there are 43 options and just one is correct. In any case, it would be
$$\frac{6\cdot 43}{\binom{49}{6}}\cdot\frac{1}{43} = \frac{6}{\binom{49}{6}}.$$
- This will happen when you have 5 out of 6 correct numbers and the other is chosen between the remaining 42 (excluding the complementary). Hence, the number of possibilities is $\frac{6\cdot 42}{\binom{49}{6}} = \frac{252}{\binom{49}{6}}$.
Alternatively, proceding as before, we would have
$$\frac{6\cdot 43}{\binom{49}{6}}\cdot\frac{42}{43} = \frac{252}{\binom{49}{6}}.$$
- The probability of failing n times is
$$\left(1-\frac{1}{\binom{49}{6}}\right)^n.$$
We want this number to be smaller than 0.5. Hence,
$$n = \frac{\log(0.5)}{\log\left(1-\frac{720}{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44}\right)} \simeq 9692843.$$

Question 2.

- The probability that the sum is ≤ 6 is $15/36$, so this is the desired value (for the sake of completeness, the second urn is chosen with probability $6/36$ and the third one with probability $15/36$).
- The answer is just (using the formula of total probabilities)
$$\frac{15}{36}\cdot\frac{3}{5}+\frac{6}{36}\cdot\frac{2}{5}+\frac{15}{36}\cdot\frac{1}{5} = \frac{72}{180} = \frac{2}{5}.$$
- We use the formula of conditioned probabilities. For that, the probability that is is both white and from urn 2 is $6/36\cdot 2/5 = 1/15$. Then, since we are conditioning by the fact of being white, we get
$$\frac{1/15}{2/5} = \frac{1}{6}.$$

Question 3.

- First of all, probabilities must add up one. Hence, $t^2 + 3t/2 = 1$, and this is a second degree equation whose solutions are $1/2$ and -2. -2 is not admissible since with that we get negative probabilities. The only possible answer is $t = 1/2$, for which you get

$$\Pr(X = 6) = 1/4, \quad \Pr(X = 10) = 1/2, \quad \Pr(X = 12) = 1/4.$$

- The expectation is

$$E(X) = \frac{1}{4} \cdot 6 + \frac{1}{2} \cdot 10 + \frac{1}{4} \cdot 12 = 9.5$$

- The variance is

$$\mathrm{Var}(X) = \frac{1}{4} \cdot (-3.5)^2 + \frac{1}{2} \cdot (10 - 9.5)^2 + \frac{1}{4} \cdot (12 - 9.5)^2 = 4.75$$

EXAM 2: DESCRIPTIVE STATISTICS

21th April 2017

Question 1. The following table represents the distribution of weights in a class of 40 students:

Weight	**Frequency**
35.5-42.5	2
42.5-49.5	11
49.5-56.5	13
56.3-63.5	9
63.5-70.5	3
70.5-77.5	2

- Determine the mean, the standard deviation and the variation coefficient.
- Estimate the median, Q_3 and p_{95}.
- Estimate to which percentile correspond the measures of 40 kg and 60 kg.

Question 2. We have studied the mistakes done by a group of students in two exams, one of Russian Literature (this will be the x variable) and the other of Number Theory (y variable). We have obtained the following table:

x_i y_i	0	1	2
0	24	6	1
1	11	19	2
2	7	8	6

- Determine the mean number of mistakes in the exam of Russian Literature, as well as the standard deviation. Do the same for the exam in Number Theory.
- Find the covariance and correlation coefficient of the two variables.

Question 3. Answer **two** of the following three questions.

- Check that
$$f(x) = \begin{cases} x/2 - 1 & \text{if } 2 \leq x < 4 \\ 0 & \text{elsewhere} \end{cases}$$
is a probability density function (for a certain random variable X). Plot it and find $P[X \leq 3]$.
- If $X \sim N(88, 6)$, what is the probability that $X < 80$? That is, find $P[X < 80]$. Determine also $P[80 < X \leq 100]$.
- If we know that 7% of the population suffers a certain pathology, determine the probability that in an enterprise of 800 workers the pathology affects to at least 80 people.

Question 1. For some parts of this exercise we will use Excel, to ease the computations.

- The mean is 54.05 and the standard deviation, 8.36. The variation coefficient is just the quotient of the two values, namely 0.155.
- For estimating the median, we must compute first the table of cumulative proportions. That way, we see that the percentage of people below 49.5 is 0.325 and those below 56.5 represent 0.65. We can do now a line joining $(49.5, 0.325)$ and $(56.5, 0.65)$. That way, we will have
$$y = 0.325 + \frac{0.65 - 0.325}{56.5 - 49.5}(x - 49.5),$$
and substituting $y = 0.5$ we get that $x = 53.27$.
Doing the same for Q_3 we get a value of 59.61. For p_{95} we just observe that this is what corresponds to 70.5.
- In this part we must proceed analogously:
$$y = \frac{0.05}{7}(x - 35.5),$$
and now putting $x = 40$ we will get that this value lies in p_3. In the same way, the value of 60 kg lies in p_{76}.

Question 2. This question is solved in the Excel file.

- For Russian Literature, the mean is 0.61 and the standard deviation, 0.67. For Number Theory, the mean is 0.88 and the standard deviation, 0.78.
- For guessing the covariance, we observe that the mean of the corresponding products is 0.75. Hence, the covariance is 0.215 and the coefficient of correlation, 0.411.

Question 3.

- To verify that $f(x)$ is a probability density function we need to check that it is non-negative (that is true because $x/2 \geq 1$ provided that $x \geq 2$) and that the area under the curve $y = x/2 - 1$ between 2 and 4 is one; but this is true since this corresponds to a triangle whose base length is 2 and height 1. For the probability that $X \leq 3$, we get again a triangle, now of base 1 and of height 1/2. Hence, $P[X \leq 3] = 1/4$.
- Standardizing the $N(88, 6)$ by considering $Z \sim N(0, 1)$, we get that
$$P(X < 80) = P\Big(Z < \frac{80 - 88}{6}\Big) = P(Z < -4/3)$$
$$= 1 - P(Z \leq 4/3) = 1 - 0.9088 = 0.0912.$$
Analogously
$$P[80 < X \leq 100] = P(-4/3 < Z \leq 2) = P(Z \leq 2) - P(Z < -4/3)$$
$$= 0.9772 - 0.0912 = 0.886.$$

- We must use the central limit theorem and approximate a binomial distribution of parameters $n = 800$ and $p = 0.07$ by a normal distribution of parameters $\mu = 800 \cdot 0.07 = 56$ and $\sigma = \sqrt{800 \cdot 0.07 \cdot 0.93} = 7.22$. We see that $\frac{79.5-56}{7.22} = 3.25$ (we take 79.5 because we are approximating a discrete variable by a continuous one), and hence the probability that the number of ill people is greater or equal than 80 is $P(Z \geq 3.25) = 0.0006$.

FINAL EXAM: BASICS OF STATISTICS

28th April 2017

In this exam, question 1 will receive a maximum score of 4 points. The other three questions will receive a maximum score of 2 points. Remember that this exam represents a 40% of the course final qualification.

Try to be as clear as possible, explaining the procedures you follow to derive your results and stating the theoretical results you are using in each moment. Moreover, in the more theoretical questions, detailed and precise descriptions are required.

Along the weekend, I will send you a message with the grade of the exam, as well as the final mark of this course.

Question 1. Each item represents 1 point of the total mark.

- We have a deck with 52 cards (13 spades, 13 hearts, 13 diamonds and 13 clovers). We take three cards from the deck. Compute the probability of obtaining 2 spades and 1 heart.
- The following table shows the height distribution in a class of 50 students.

Height	**Frequency**
140-150	3
150-160	7
160-170	20
170-180	18
180-190	2

  Estimate the **mean** and the **median** of the height.

- Explain in all detail the meaning of the following words: **percentile**, **probability density function**, **type I error** and **test statistic**.
- Discuss in detail what is the usefulness of the chi-squared tests. Give examples in which it is used and discuss how you would carry out one in with whole detail, explaining what is the null hypothesis, the alternative hypothesis, the test statistic, the significance level and the rejected region.

Question 2. (2 points) As in "Exam 1", we have three urns with the following compositions (W for white and B for black):

$$U_1(7W, 3B), \quad U_2(6W, 4B), \quad U_3(5W, 5B).$$

A fake dice is rolled. In this fake dice, the probability of obtaining $1, 2, 3, 4$ or 5 is equal (call this number p), and this value is half of the probability of obtaining 6.

- Determine the probability of obtaining 6.

Then, if the resulting number is even we choose U_1, if it is one we choose U_2 and elsewhere we choose U_3. Then, two balls are taken from the chosen urn.

- Determine the probability that both balls are white.
- If we know that we have obtained two white balls, what is the probability that the outcome obtained when rolling the dice was 1?

Question 3. (2 points) The following table shows the earning (in million dollars) of different companies as a function of the number of employees they have.

Employees	**Earnings**
45	11
54	15
37	11
41	13
35	11
29	7
61	18
45	14
43	11
49	16
34	11
40	13

- Calculate the correlation coefficient.
- Fit the data to a line using the method of linear regression.
- Use the previous results to predict the earnings when we have 35 employees, and the number of employees when the earnings are 20 million dollars.

Question 4. (2 points) A test to determine the average height of a set of chairs was performed and we have obtained the following results (in centimeters):

$$26.7; 25.8; 24.0; 24.9; 26.4; 25.9; 24.4; 21.7;$$

$$24.1; 25.9; 27.3; 26.9; 27.3; 24.8; 23.6; 25.0.$$

Assuming that height of chairs follows a normal distribution, find 90% and 95% confidence intervals for average height in the following two cases:

- Having the previous knowledge that $\sigma = 1.5$.
- Having no knowledge about the value of the standard deviation.

Question 1. We present a sketch of how the solutions should go:

- Let us find the probability of obtaining first spade, secondly space and lastly heart. This is just
$$\frac{13}{52} \cdot \frac{12}{51} \cdot \frac{13}{50} = \frac{13}{850}.$$
Then, this result should be multiplied by 3 (heart can go in any of the three positions). Finally, we get $\frac{39}{850}$.

- One possibility for estimating the mean is considering that all the measure in a certain range are equal to the central value $(145, 155, \ldots, 185)$. That way, we can see (with Excel for instance) that the mean would be 166.8.
For the median we study the cumulative proportions. A 20% of the population is ≤ 160 and a 60% is ≤ 170. Then, if we draw a line joining $(160, 20)$ and $(170, 60)$ we will obtain that the point corresponding to 50% is 167.5.

- **Percentiles:** the percentile p_i is the value that leaves below it the i% of the population, and above it, the $(100 - i)$%. Although there is no universal convention, typically, in an ordered set, for the p_i percentile in a set of n, we shall consider the value that occupies the first position greater or equal than $n \cdot \frac{i}{100}$.
Probability density function: for a continuous random variable X, the probability density function $f : \mathbb{R} \to \mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,
$$P[a < X \leq b] = \int_a^b f(x)\, dx$$
and it satisfies the following two properties:
  - $f(x) \geq 0$ for all $x \in \mathbb{R}$.
  - The area under the curve is 1, that is, $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

  Roughly speaking, the probability of being between a and b is the area under the curve.
Type I error: when doing hypothesis testing, is the error that occurs when H_0 is rejected being true. This is controlled by the parameter α, that measures the probability of doing one such mistake.
Test statistic: is a function of the sample data on which the decision of reject H_0 or do not reject H_0 is to be based on.

- The aim in this type of tests is to compare proportions of two or more populations. The tests typically carried out are those of goodness of fit and independence. Suppose we perform an experiment such that their results can be classified into k categories of cells and that it is repeated n times. Furthermore, assume that the odds or proportions of the different results are $p_1, \ldots, p_k$, with $\sum p_i = 1$ and that in the total of the n repetitions the frequencies observed were $O_1, \ldots, O_n$ with $\sum O_i = n$.
Then, the statistical test is
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i},$$
where $e_i = np_i$ is the expected frequency. We reject H_0 (the hypothesis that $\pi_1 = p_{10}, \ldots, \pi_k = p_{k0}$) when $\chi^2 > \chi^2_{\alpha,k-1}$, being α the significance level we fix beforehand. Observe that the null hypothesis is that at least one of the $p_i \neq p_{i0}$.

Question 2.

- The probability of obtaining 6 is $2p$. Then, $5p + 2p = 1$ and hence $p = 1/7$.
- Let us first guess the probability that both balls are white and from urn 1 is $\frac{4}{7} \cdot \frac{7}{10} \cdot \frac{6}{9} = \frac{4}{15}$.
  Similarly, the probability that both balls are white and from urn 2 is $\frac{1}{7} \cdot \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{21}$.
  Finally, both balls are white and from urn 3 with a probability of: $\frac{2}{7} \cdot \frac{5}{10} \cdot \frac{4}{9} = \frac{4}{63}$.
  Hence, the probability that both are white is

  $$\frac{4}{15} + \frac{1}{21} + \frac{4}{63} = \frac{84 + 15 + 20}{315} = \frac{119}{315}.$$

- Using Bayes' formula, the answer is just

  $$\frac{15/315}{119/315} = \frac{15}{119}.$$

Question 3. This question can be solved using Excel.

- The covariance is 21.396 and the correlation coefficient, 0.899. This means a positive correlation (quite strong): when we hire more employees, we earn more money.
- Using the formulas we have learned, we get that the equation of the fitted line is

  $$y = 0.292x + 0.086.$$

- Using the previous line, we see that when we have 35 employees the expected earnings are 10.32 million dollars, and when we earn 20 million dollars is because we have $68.12 \sim 68$ employees.

Question 4. First of all, we see with Excel that the mean is 25.29 and the sample standard deviation is 1.527.

- In this first case, using the formulas for the confidence intervals, we get

  $$\left(25.29 - z_{1-\alpha/2}\frac{1.5}{\sqrt{16}}, 25.29 + z_{1-\alpha/2}\frac{1.5}{\sqrt{16}}\right).$$

  When we want a confidence interval of 90%, $\alpha = 0.1$ and we have to look in the normal table for $z_{0.95} = 1.645$; when we want 95%, $\alpha = 0.05$ and we have that $z_{0.975} = 1.96$.

- If we must estimate the standard deviation (we use as usual the sample standard deviation), the confidence interval is related with a certain t-value, since the number of observations is small and we cannot approximate it with the normal:

  $$\left(25.29 - t_{15,1-\alpha/2}\frac{1.527}{\sqrt{16}}, 25.29 + t_{15,1-\alpha/2}\frac{1.527}{\sqrt{16}}\right).$$

  The values we obtain with the table are $t_{15,0.95} = 1.753$ and $t_{15,0.975} = 2.131$.

QUIZZ 1: COUNTING AND PROBABILITY

12th April 2017

Choose one of the following questions and answer it.

Question 1. We roll four (fair) dices in a row. Consider the following events:

$$A = \{\text{the sum of the four dices is } 6\}.$$

$$B = \{\text{the sum of the first two dices is } 3\}.$$

- Find $\Pr(A)$ and $\Pr(B)$ (2 points each).
- Find $\Pr(A \cap B)$ (2 points).
- Knowing that the sum of the first two dices is 3, what is the probability that the sum of the four dices is 6? (2 points)
- Knowing that the sum of the four dices is 6, what is the probability that the sum of the two first dices is 3? (2 points)

Question 2. A certain (evil) professor teaches a course in Harbour Space for 10 students, 5 boys and 5 girls.

- Determine in how many ways he can pass to at most 3 students (2.5 points).
- Determine in how many ways he can pass to 2 boys and 1 girl (2.5 points).
- Determine in how many ways he can pass to at most 2 boys and exactly 1 girl (2.5 points).
- Now, the professor wants to give 2 passes: one with letter A and one with letter B. In how many ways can he do this if there must be one pass from each gender? (2.5 points).

Question 1.

- There are 10 possible ways of getting a sum of 6:

  $$3+1+1+1, \quad 1+3+1+1, \quad 1+1+3+1, \quad 1+1+1+3,$$

  $$2{+}2{+}1{+}1, \quad 2{+}1{+}2{+}1, \quad 2{+}1{+}1{+}2, \quad 1{+}2{+}2{+}1, \quad 1{+}2{+}1{+}2, \quad 1{+}1{+}2{+}2.$$

  Observe that once we get that the possibilities are of the form $3+1+1+1$ and $2+2+1+1$ (with the corresponding permutations) we can deduce that there are 4 options for the first form, because $\frac{4!}{3!1!1!1!} = 4$, and 6 for the second, since $\frac{4!}{2!2!1!1!} = 6$. There are 1296 options, so $\Pr(A) = 10/1296 = 5/648$.
  On the other hand, if the sum of the first two dices is 3, it must be either because I have $1+2$ or $2+1$. We get $\Pr(B) = 2/36 = 1/18$.

- In this case, we have that the sum of the first two dices is 3 and the sum of the last two is also 3. We have two possibilities for getting a sum of 3 in the first two, and also two possibilities for a sum of 3 in the last two. Then, there are $2 \cdot 2 = 4$ options, and since the total number of possibilities is 1296, we get $4/1296 = 1/324$.

- One option for solving the question is just observing that

  $$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{4/1296}{2/36} = \frac{2}{36} = \frac{1}{18}.$$

  Alternatively, we just want the sum of the last two dices to be 3. This can be done in two ways out of 36.

- Again, by conditioned probability

  $$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{4/1296}{10/1296} = \frac{4}{10} = \frac{2}{5}.$$

  Alternatively, when we have listed the ten possibilities for A, we see that just 4 satisfy that the sum of the first two dices is 3.

Question 2.

- This corresponds to choosing a subset of 3, 2, 1 or 0 students, in a class of 10 people. This can be done in

  $$\binom{10}{3} + \binom{10}{2} + \binom{10}{1} + \binom{10}{0} = 120 + 45 + 10 + 1 = 176 \text{ ways.}$$

- There are $\binom{5}{2} = 10$ ways to pick up two boy, and 5 ways to pick up a girl. Hence, there are 50 possible choices.

- There are $\binom{5}{2} + \binom{5}{1} + \binom{5}{0} = 16$ ways to pass to at most two boys, and 5 ways to pass to exactly one girl. Hence, the number of choices is 80.

- First of all, we have to select a pair formed by one girl and one boy, and this can be done in $5 \cdot 5 = 25$ ways; once the pair is chosen, there are two ways of distributing the marks between them, and hence we have 50 possibilities.

QUIZZ 2: DESCRIPTIVE STATISTICS

19th April 2017

Definitions. Explain with your own words the meaning of the following words, discussing its usefulness in statistics (1 point each): **trimmed arithmetic mean**, **variation coefficient**, **cumulative frequency**.

Question 1. The following data correspond to the residual chlorine in the water tank of a city at various times after water have been treated with chemicals to keep it suitable for human consumption:

Time (hours)	**Cl (ppm)**
2	1.8
4	1.5
6	1.4
8	1.1
10	1.1
12	0.9

- Calculate the sample correlation coefficient and comment on the result (2.5 points).
- Get a fitted line to predict the residual chlorine in terms of time since the water was treated with chemicals (2.5 points).
- Use the previous line to estimate the residual chlorine in the tank 7 hours after it has been treated (2 points).

Definitions.

- **Trimmed arithmetic mean:** is that computed by first ordering the data values from smallest to largest, deleting a selected number of values from each end of the ordered list and finally averaging the remaining values. The trimming percentage is the percentage of values deleted from each end. The $\alpha\%$ trimmed mean is the arithmetic mean of data remaining in the sample after removing the $\alpha/2\%$ of the largest and the $\alpha/2\%$ of the smallest scores. This is done because arithmetic mean is very sensitive to extreme values and we want to avoid this phenomenon.

- **Variation coefficient:** it is the quotient $\frac{\sigma}{\mu}$, one way of "normalizing" the standard deviation, since in many situations we will be interested in its size in relation with the mean. A value of $\sigma = 1$ when $\mu = 10^9$ is usually acceptable, but when $\mu = 2$, this same value of sigma is often very bad.

- **Cumulative frequency:** in a frequency table, when we have ordered data grouped in classes, the (absolute) cumulative frequency is the number of observations in our sample that are below or at the same level to a certain value. Alternatively, it is the resulting sum of the sizes of all the classes that are below or at the same level than the class we are considering. We sometimes talk about relative cumulative frequency (or cumulative proportion) for that same concept but now normalizing by the size of the sample.

Question 1. In the Excel file we have calculated the different parameters of this bivariate distribution. In particular, if x refers to the time and y to the concentration, we have the following:

- $\bar{x} = 7$.
- $\bar{y} = 1.3$.
- $\sigma_x = 3.42$.
- $\sigma_y = 0.3$.
- $\sigma_{xy} = -1$.
- $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = -0.9759$.

In particular, for the first part, we know that the correlation coefficient is -0.9759, and this indicates that there is a strong negative correlation between the variables: the greater the time, the lower the concentration of residual chlorine is.
For the regression line, the slope is just $m = \frac{\sigma_{xy}}{\sigma_x^2} = -0.086$; on the other hand, the value of y at the origin will be $n = \bar{y} - m\bar{x} = 1.9$. All in all, the equation of the line is

$$y = -0.0857x + 1.9.$$

Finally, substituting $x = 7$ we get $y = 1.3$.

QUIZ 3: INTERVAL ESTIMATION

25th April 2017

Definitions. Explain with your own words the meaning of the following words, discussing its usefulness in statistics (1 point each): **sampling**, **estimator**, **confidence interval**.

Now, choose one of the following two questions (7 points):

Question 1. In order to estimate the mean weight of 15-year old children in a given country, we select a random sample of 100 people. We obtain the following values: $\bar{x} = 52.5$ kg and $s = 5.3$ kg. We do the following affirmation: "the mean weight of 15-year old children in the country is between 51 and 54 kg". With which level of probability are we making our affirmation?

Question 2. A teacher wants to estimate the mean height of all his students with an error smaller than 0.5 cm using a sample of 30 children. Knowing that $\sigma = 5.3$ cm, determine the confidence level.

Definitions. We present here a summary of what was expected.
Sampling: is the process by which we extract data from a population by selecting a sample, the group that will be considered for our study and that is not the whole population (that alternative study will be via a census). There are several types of sampling, as the systematic one, the stratified one or the one by clusters, and according to the situation one may be more adequate that others.
Estimator: is a function of the sample appropriate to estimate a population parameter. For example, when we have a normal distribution, an estimator for the mean is just the sample mean of the observations.
Confidence interval: a confidence interval corresponding to a probability p is an interval in which our parameter is found with total probability p. Typically, we are concerned about intervals for the mean, that we consider that are centered in the estimated value.

Question 1. In this problem we are concerned about estimation of intervals for the mean for unknown variance; however, since n is big, we can use the Z-table instead the t-tables. In particular, we know that the mean will follow a distribution

$$X \sim N\Big(52.5, \frac{5.3}{\sqrt{100}}\Big) = N(52.5, 0.53).$$

Hence,

$$P(X > 54) = P\Big(\frac{54-52.5}{0.53}\Big) = P(Z > 2.83) = 1 - 0.9977 = 0.0023.$$

Analogously, $P(X < 51) = 0.0023$ and hence the probability of belonging to the interval is 0.9954, that is, 99.54%.

Question 2. The error is given by

$$E = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

and then

$$z_{1-\alpha/2} = \frac{E\sqrt{n}}{\sigma} = 1.634.$$

Then, $1 - \alpha/2 = 0.9488$ and consequently $\alpha = 0.1024$ and the probability of being inside the interval is 89.76.

QUIZ 4: HYPOTHESIS TESTING

27th April 2017

Definitions. Explain with your own words the meaning of the following words, discussing its usefulness in statistics (1 point each): **null hypothesis**, **significance level**, **rejected region**.

Question 1. A mall sells TV, DVD and digital cameras (DC). To determine whether there is a relationship between the method of payment and purchased product sales in the last month, sales and method of payment were recorded.

	TV	**DVD**	**DC**	**Total**
Cash	11	4	7	22
Card	52	19	12	83
Credit	27	32	11	70
Total	90	55	30	175

Determine if it can be concluded to 10% of significance level that there exists relationship between the purchased product and how to pay for it (7 points).