FIRST HAND IN: A REVIEW OF COUNTING

10th April 2017

During the first class, we have presented the course and we have insisted on the importance of a good knowledge of probability to face statistical problems. This first hand-in will take care of some of the materials covered in the previous course, focusing on counting procedures. Each of the following questions will receive a mark between 0 and 2 points (yes, the sum is 12, but some of them are not easy!):

1. Explain with your own words the concepts of permutation (with and without repetition) and combination (with and without repetition). Give examples.
2. With the letters $\{M, N, O, P, Q, R, S\}$ we want to form a 4-letter word (no repetitions allowed). Determine in how many ways we can do that.
3. Determine the number of words in the alphabet $\{a, b, c, d\}$ that use exactly three times letter a, 5 times letter b, twice letter c and once letter d.
4. An enterprise of detectives have 12 microphones that must hide in the offices of 5 people, say A, B, C, D, E. Find in how many ways this can be done, taken into account that in the offices of both A and B they must put at least 3 microphones and in the others at least one.
5. A different enterprise must infiltrate people between the staff of three political parties, X, Y, Z. Determine in how many ways the 7 agents $a_1, a_2, \ldots, a_7$ can do the task, assuming that each party has at least one infiltrated agent?
6. Give a combinatorial proof of the following identities:

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n,$$

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

1. This is written in the class notes.

2. This corresponds to permutations (the order matters) without repetition. Consequently, this can be done in
$$7 \cdot 6 \cdot 5 \cdot 4 = 840$$
ways.
In the version I gave you, the letters were $\{M, M, M, N, N, O, P\}$ and we wanted a word of seven letters. We must choose the three places where M will go; this can be done in $\binom{7}{3} = 35$ ways; once we do this, we must choose two places (between the four that are free) for the N. This can be done in $\binom{4}{2} = 6$ ways. Finally, two places are free to put the O and the P, so this adds a factor of 2. The final answers is
$$35 \cdot 6 \cdot 2 = 420.$$

3. We must choose the three places where a will go; this can be done in $\binom{11}{3} = 35$ ways; once we do this, we must choose five places (between the eight that are free) for the b. This can be done in $\binom{8}{5} =$ ways. Finally, two places must be for the c, so we have a factor of $\binom{3}{2}$. The final answers is
$$\binom{11}{3} \cdot \binom{8}{5} \cdot \binom{3}{2} = \frac{11!}{3!5!2!1!} = 27720.$$

4. First of all, I put the "compulsory" microphones, and then I have 3 remaining microphones, that I must distribute in 5 offices. This corresponds to combinations (the order of the microphones does not matter) with repetition, so the answer must be
$$\binom{3+5-1}{3} = 35.$$
Alternatively, we can proceed as follows: we can put either the three in the same offices (and this can be done in 5 ways); either two in one office and one in a different one (this is $5 \times 4 = 20$); or we can put three microphones in three different offices (this corresponds to $\binom{5}{3}$, since order does not matter). Again, you get $5 + 20 + 10 = 35$ ways.

5. The number of ways of distributing the 7 agents without any restriction is 3^7 (permutations with repetitions). We must remove those that leave either X, Y or Z unattended. X is unattended when agents only take care of Y and Z (there are $2^7 = 128$ ways to do this); Y is unattended when agents only take care of X and Z (there are $2^7 = 128$ ways to do this) and finally Z is unattended when agents only take care of X and Y (there are $2^7 = 128$ ways to do this). This would give a result of $2187 - 3 \cdot 128$, but those times in which two parties are unattended have been discounted twice, so they must be added once; if exactly two parties are unattended this means that all the agents take care of the same party, and this can be done in three ways (since there are three parties). The final result is hence
$$2187 - 384 + 3 = 1806.$$

6. For the first part, $\binom{n}{k}$ is the number of ways of selecting k students to pass in a class of n; this can be done in two ways: either the first student passes and

then you pass $k-1$ other students, that can be done in $\binom{n-1}{k-1}$ ways, or the first one fails and you must pass $k-1$ students between the remaining ones, in $\binom{n-1}{k}$ possible ways. All in all,

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

For the second part, we continue with our class of n students. There are 2^n ways of distributing the final mark according to “pass” of “fail” (there are two possibilities for each student). Alternatively, we may fix first the number of passes, that is a number between 0 and n, say k. Once k is fixed, there are $\binom{n}{k}$ ways of selecting the k students to pass. These two numbers must be the same, and hence

$$2^n = \sum_{k=0}^{n} \binom{n}{k}.$$

SECOND HAND IN: AN INVITATION TO PROBABILITY

11th April 2017

During this second class, we have reviewed the definitions of probability space, independence and conditioned probability. At the end, we have discussed Bayes' formula, one of the main tools for a basic study of probability problems. In this sheet, I present you now three problems. Each one will receive a grade between 0 and 4 points (although the maximum qualification will be 10).

1. A certain partner has 5 children. Determine the probability of having at least one boy and at least one girl (1 point).
   How many children must they have if they want this probability to be higher than 99% (1.5 points).
   Repeat the calculation assuming that the probability that the child is a boy is 2/3, and therefore, that the probability that the child is a girl is 1/3 (1.5 points).

2. We roll two (fair) dices. Let A be the event that the resulting sum is even. Let B the event that the resulting sum is ≥ 10. Find $\Pr(A)$ (1 point) and $\Pr(B)$ (1 point). Are these events independent? (0.5 points)
   Determine the probability that when rolling three dices the sum is less or equal than 5 (1.5 points).

3. We toss 10 (fair) coins. Determine the probability that both the first and the second one are head (0.5 points).
   Determine the probability of obtaining exactly 8 heads (1 points).
   Determine the probability of obtaining exactly 8 heads knowing that the first and the second one are head (1.5 points).
   Determine the probability that the first and the second one are head knowing that there are exactly 8 heads (1 points).

1. The probability that the five children were boys is $(1/2)^5 = 1/32$; the same holds for girls. Consequently, the probability of having at least one boy and at least one girl is $1 - 1/32 - 1/32 = 15/16$.

   If we repeat the calculation with n children, the probability will be $(2^{n-1} - 1)/2^n$. We want this value to be greater than 0.99, that is, $2^{n-1} - 1 > 0.99 \cdot 2^{n-1}$, or what is the same $2^{n-1} > 100$. Now, taking logarithms (or just observing that the first power of 2 above 100 is 128) we conclude that $n = 8$ is the smallest possible value.

   For the last question, we realise that the probability that the five children were boys is $(2/3)^5 = 32/243$; for girls, the answer is $1/243$. Consequently, the probability of having at least one of each gender is $1-33/243 = 70/81$. For an arbitrary n, we want $(2^n + 1)/3^n$ to be small than 0.01 (since we want the complementary to be bigger than 0.99). We can estimate the value of n using logarithms (since the 1 of the numerator is negligible), and then check that the answer is correct, or just give values to n. In any case, we see that $n = 12$ is the first value that makes the quantity smaller than 0.01.

2. We can either check the probability of the sum being $2, 4, 6, 8, 10, 12$ and then add the values, or alternatively observe that the sum is even either if both results are even (this happens with probability $1/2 \cdot 1/2 = 1/4$) or if both are odd (by the same reason, probability $1/4$). We conclude that the answer is $\Pr(A) = 1/2$. On the other hand, the sum is ≥ 10 if it is either 10 (probability $3/36$), 11 (probability $2/36$) or 12 (probability $1/36$). Hence, $\Pr(B) = 1/6$. We finally see that the intersection is formed by two elementary events: sum 10 and sum 12. Then, $\Pr(A \cap B) = 1/9$ that is different from $1/4 \cdot 1/6 = 1/24$.

   The sum of three dices will be 3 just in the case $1 + 1 + 1$ (probability $1/216$); there are three cases of sum 4 (probability $3/216$); finally, there are six cases of sum 5 (probability $6/216$). Hence, the answer is $5/108$.

3. The probability that both the first and the second are head is just $1/2 \cdot 1/2 = 1/4$. For that of obtaining 8 heads, I have to select those places where heads will go; this can be done in $\binom{10}{8} = 45$ ways. Hence, the probability is $45/1024$.

   If we know that the first and the second are head, we just want 6 heads out of the following eight. The probability, by the same reason as before, is $\binom{8}{2}/256 = 7/64$. Alternatively, using conditioned probabilities, we see that the probability of the intersection is $28/1024$ (we have as many favourable cases as ways of choosing 6 places in a set of 8), and dividing the result by $1/4$ we reach the same result.

   For the last question, once we know the probability of the intersection, by conditioned probability, we have $\frac{28/1024}{45/1024} = 28/45$.

THIRD HAND IN: MORE ON PROBABILITY

12th April 2017

During this third class, we have reviewed the concepts of previous classes and introduce the concepts of random variable (as well as that of expectation and variance). In this sheet, I present you now three problems. Each one will receive a grade between 0 and 4 points (although the maximum qualification will be 10).

1. We have two dices, one is fair and the other is fake. The fake one has 4 times number one and 2 times number two. We choose a dice at random and roll it twice. Determine the probability of obtaining 1 the first time and 2 the second time (2 points).
   If we know that the result of the first time was 1 and that of the second time was 2, determine the probability of having chosen the fair dice (2 points).

2. A certain random variable X takes 3 different values, $1, 2, 3$, with the following probability function

   $$\Pr(X=1) = 3k-2, \quad \Pr(X=2) = 3-4k, \quad \Pr(X=3) = k,$$

   where k is a certain real number. Find the range of values of k such that what we have defined is really a probability function. For instance, $k=0$ is not possible since this would imply $\Pr(X=1) = -2$. (1.5 points).
   Find the expectation and the variance of X (2 points).
   What is the mode, that is, the value that X takes with higher probability? (0.5 points).

3. We have an enterprise of ten taxis. Each taxi does not work a given day with a probability of 1/5. Determine the probability of having strictly more than two broken taxis a given day (2 points).
   Is it more or less likely to have all the taxis working properly or to have exactly three broken taxis? (0.5 points).
   Each broken taxis costs us 500 euros a day. Determine how many euros we lose in mean every day? (1.5 points).

1. For the fair one, the probability is just $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$; for the fake one, you get $\frac{4}{6} \cdot \frac{2}{6} = \frac{8}{36}$. Hence, since it is equaly likely to choose one dice or the other, the probability is
$$\frac{1}{2} \cdot \frac{1}{36} + \frac{1}{2} \cdot \frac{8}{36} = \frac{9}{72} = \frac{1}{8}.$$
Concerning the second question, we use Bayes' formula (conditioned probability), taken into account that the probability of choosing the first dice and obtaining 1 and 2 at the same time is $\frac{1}{2} \cdot \frac{1}{36} = \frac{1}{72}$. Hence, we get $\frac{1/72}{1/8} = 1/9$.

2. Two things must be satisfied: probabilities must add up one and each probability must be in the interval $[0, 1]$. For the sum being one, observe that
$$\Pr(X = 1+) + \Pr(X = 2) + \Pr(X = 3) = 3k - 2 + 3 - 4k + k = 1,$$
no matter which value k takes. The other conditions are
$$0 \leq 3k - 2 \leq 1 \Leftrightarrow k \geq 2/3 \text{ and } k \leq 1,$$
$$0 \leq 3 - 4k \leq 1 \Leftrightarrow k \geq 1/2 \text{ and } k \leq 3/4,$$
$$0 \leq k \leq 1 \Leftrightarrow k \geq 0 \text{ and } k \leq 1.$$
Then, we get that $k \geq 2/3$ and $k \leq 3/4$. We conclude that $k \in [2/3, 3/4]$.
For the expectation,
$$E(X) = 1 \cdot \Pr(X = 1) + 2 \cdot \Pr(X = 2) + 3 \cdot \Pr(X = 3) = 3k - 2 + 2(3 - 4k) + 3k = -2k + 4.$$
For the variance, we must compute the following quantities:
$$(1 - E(X))^2 = (1 - (-2k + 4))^2 = (2k - 3)^2 = 4k^2 - 12k + 9,$$
$$(2 - E(X))^2 = (2 - (-2k + 4))^2 = (2k - 2)^2 = 4k^2 - 8k + 4,$$
$$(3 - E(X))^2 = (3 - (-2k + 4))^2 = (2k - 1)^2 = 4k^2 - 4k + 1.$$
Now, by definition
$$\mathrm{Var}(X) = \Pr(X = 1) \cdot (1 - E(X))^2) + \Pr(X = 2) \cdot (2 - E(X))^2 + \Pr(X = 3) \cdot (3 - E(X))^2$$
$$= (3k - 2)(4k^2 - 12k + 9) + (3 - 4k)(4k^2 - 8k + 4) + k(4k^2 - 4k + 1) = -4k^2 + 12k - 6.$$
In the range of values we are working with, the values are positive and the sum is 1; since $\Pr(X = 3) > 2/3$ the other two values must be smaller than $1 - 2/3 = 1/3$, so the bigger one is always $X = 3$. Hence, the mode is 3.

3. The probability of having all taxis working is $(4/5)^{10}$. That of having exactly one broken taxi is $10(1/5)(4/5)^9$, and that of having exactly two is $45(1/5)^2(4/5)^8$, since $45 = \binom{10}{2}$. Adding up these two values you get $\simeq 0.678$. Since we are interested in the complementary event, we get 0.322.
The probability that all taxis work is $(4/5)^{10} \simeq 0.107$. The probability that exactly three taxis are broken is $\binom{10}{3}(0.8)^7(0.2)^3 \simeq 0.201$, and this last number is bigger.
Let X_1 be the random variable that takes the value 1 if the taxi is broken and 0 elsewhere. We clearly have $E(X_1) = 1/5$. Define in the same way $X_2, X_3, \ldots, X_{10}$. Then, we observe that the expected number of broken taxis is
$$E(X_1 + X_2 + \ldots + X_{10}) = E(X_1) + E(X_2) + \ldots + E(X_{10}) = 10 \cdot 1/5 = 2.$$
Hence, the expected losses are $500 \cdot 2 = 1000$ euros.

FORTH HAND IN: WALKING TOWARDS STATISTICS

17th April 2017

In this class (after a not very difficult exam), we have started to introduce some terminology related with statistics. Since we do not still have a lot of machinery for solving problems, we present very easy exercises just for you to get used to the treatment of data. On the following days, we will go deeper into some of these features.

1. We have measured the heart rate per minute of 30 people. We have obtained the following results:

$$87, 85, 61, 51, 64, 75, 80, 70, 69, 92,$$
$$80, 79, 82, 74, 90, 76, 72, 73, 63, 65,$$
$$67, 71, 88, 76, 68, 73, 70, 76, 71, 86.$$

   Find the mean and the standard deviation of this data (1 point).
   Which percentage lies in the interval $(\bar{x} - \sigma, \bar{x} + \sigma)$? (1 point).
   Represent graphically the distribution grouping the data in intervals (1 point).
   Explain how would you determine $\bar{x}$ and σ with the grouped data and compare these results with those in the first item (1 point).

2. We have measured heights in a group of 40 people. The results we have obtained are summarized in the following table:

Height	**Frequency**
152	2
156	4
161	11
166	14
171	5
176	4

   Determine the mean (average), the standard deviation and the variation coefficient. Determine also the harmonic, geometric and quadratic mean (0.5 point each, up to a total of 3).

3. In the following distribution of grades, find the median, Q_1, Q_3, p_{80}, p_{90} and p_{99} (0.5 points each, up to a total of 3):

Grade	**Frequency**
1	7
2	15
3	41
4	52
5	104
6	69
7	26
8	13
9	19
10	14

For the first two exercices, the use of Excel is highly advisable. We will show the results we have obtained with it.

1. The mean is 74.47 and the standard deviation is 9.21. Observe that this last value can be computed either using the function Excel has or with the definitions we have learnt.
   For the second part, we must determine how many numbers are in the interval $(65.26, 83.67)$; we count that there are 19 such values, which represent a 63.33% of the total.
   Since the smallest number is 51 and the biggest is 92, we can form 9 intervals of length 5, namely $[50, 54]$, $[55, 59]$, $\ldots$, $[90, 94]$. Then, we count how many observations are in each interval and we can form a bar diagram as the one we show in the Excel file.
   Finally, we can determine the mean and the standard deviation assuming that all the data in the interval takes a value equal to the median (in this case 52, 57, $\ldots$, 92). This yields a very similar result: the mean now is 74.17 and the standard deviation 8.53.

2. Again, the use of Excel directly gives the mean (164.5), the standard deviation (6.14) and the variation coefficient (0.037).

3. Since we have 360 numbers, the median is the average of those occupying position 180 and 181. Computing the cumultative frequencies, we see that both values are 5. In the same way, $Q_1 = 4$ and $Q_3 = 6$. For the percentiles, we have that p_{80} is that values that leaves $360 \cdot \frac{80}{100} = 288$ values below it and 72 above it. Since there are exactly 288 values that are ≤ 6 and 72 that are ≥ 7, the most correct is to consider that p_{80} is 6.5, althogh according to the convention we follow we can choose either 6 or 7. p_{90} leaves 324 below and 36 above, so $p_{90} = 9$ and by the same procedure, $p_{99} = 10$.
   For the sake of completeness, we include a frequency table:

Grade	**Frequency**	**Cumulative frequency**	**Cumulative proportion**
1	7	7	0.02
2	15	22	0.06
3	41	63	0.18
4	52	115	0.32
5	104	219	0.61
6	69	288	0.80
7	26	314	0.87
8	13	327	0.91
9	19	346	0.96
10	14	360	1.00

FIFTH HAND IN: BIVARIATE DISTRIBUTIONS

18th April 2017

Once we have presented the basic definitions concerning statistics, we can pass to a deeper topics, bivariate distributions and regression lines. In this homework assignment, we begin with a problem with percentiles and then two classical exercices: in one you are requested to adjust a set of data to a certain line, while in the other you have to study correlations in a table with Excel.

1. The heights of 40 students are given in the following table:

Height	**Number of students**
158.5-163.5	1
163.5-168.5	5
168.5-173.5	11
173.5-178.5	14
178.5-183.5	6
183.5-188.5	3

   Determine the percentile corresponding to a height of 180 cm and explain what does this mean (1 point).
   Estimate the values of Q_1 and Q_3 (1 point).

2. Consider the following bidimensional distribution:

Expenses	**Sells**
1	10
2	17
3	30
4	28
5	39
6	47

   Determine the two possible regression lines (1 point each).
   Determine the corelation coefficient between x and y (1.5 points).
   Estimate the values $\hat{y}(5.5)$ and $\hat{x}(15)$ and explain their meaning (1.5 point).

3. In the Excel file "Analysis of qualifications" you will find the results of 86 people in an exam that consisted on 6 problems; each problem received a maximum score of 7 points.
   First of all, determine the average of the students, the median, the quartiles and the standard deviation. Do the same with each of the problems (2 points).
   Study the corelation between the final grade and the grade of each of the problems. Determine which problem corelates more with the final grade, as well as the corelation between the different problems (2 points).
   Which kind of graphs do you think are relevant to summarize this type of data? (1 point).

1. We need a table with the cumulative proportions:

Height	**Students**	**Cumulative frequency**	**Cumulative proportion**
158.5-163.5	1	1	0.025
163.5-168.5	5	6	0.15
168.5-173.5	11	17	0.425
173.5-178.5	14	31	0.775
178.5-183.5	6	37	0.925
183.5-188.5	3	40	1

Let us find the equation of the line that measure the cumulative proportion in the interval $[178.5, 183.5]$. We have

$$y = 0.775 + \frac{0.925 - 0.775}{183.5 - 178.5.}(x - 178.5),$$

$$y = 0.775 + 0.03(x - 178.5) = 0.03x - 4.58.$$

Hence, if $x = 180$ we get $y = 0.82$. This means that a person with a height of 180 cm is in p_{82}: 18% of the population is taller than him and 82% is smaller.
For the Q_1, we must find the equation in the interval $[168.5, 173.5]$:

$$y = 0.15 + \frac{0.425 - 0.15}{173.5 - 168.5}(x - 168.5) = 0.15 + 0.055(x - 168.5) = 0.055x - 9.1175.$$

If we put $y = 0.25$, we get $x = 170.31$.
We proceed in the same way for the Q_3 and we get $x = 178.1$.

2. If we take as the independent variable the expenses (say x) and as the dependent variable the sells (y), we will have (done in Excel) that the regression line is

$$y = 3.6 + 7.11x.$$

Now, if we change their roles, we obtain

$$x = 0.13y - 0.32$$

that can be also written isolating the y-variable and we would obtain

$$y = 7.47x + 2.37$$

that is not exactly the first equation as the first one since the correlation is not perfect. It takes the value of 0.976.
Concerning the last part, we have to estimate the value of sells knowing that expenses are 5.5. Using the first line, we obtain 42.73; the second one gives a value of 43.43. Finally, knowing that sells are 15, we are requested to estimate expenses; again, line one gives a value of 1.60 and line two a value of 1.69.

3. This exercise is summarized in the Excel file. The mean of all the students in the exam is 17.56 and the standard deviation, 11.38. Results for each problem are summarized in the corresponding tables. For the median, since we have 86 students, we take the average of the 43rd and the 44th, that gives a value of 17. Concerning the quartiles, there is not a universal convention, and one possibility

is doing the median of each half, what gives, for Q_1 a value of 8 and for Q_3 a value of 24. Again, we can do the same for each problem.

In the Excel file we have presented a table of covariances and correlations. For instance, we see that we have values ≥ 0.6 for the correlations between problems 1 and 2 and between 5 and 6. The problem that has a greater relation with the total score is problem 5, while the most related one is problem 4.

For explaining how well different problems have gone, histograms are very useful. We can also present data in some kind of bar chart to see how many people have got any of the possible scores and how these are distributed; for instance, there are many people around 17 points but few with 30 points or higher.

SIXTH HAND IN: THE NORMAL DISTRIBUTION

20th April 2017

After having done during these days a more or less complete picture of both the basics of probability theory and also of descriptive statistics, we now move to the study of continuous random variables, focusing on the normal distribution, that will allow us to study the foundations of interval estimation, hypothesis testing...

1. In a normal distribution $X \sim N(6, 4)$, find the following probabilities (1 point each):
   - $P[X \leq 3]$.
   - $P[X \geq 12]$.
   - $P[5 \leq X \leq 8]$.
2. In a distribution $X \sim \text{Bin}(200; 0.3)$, determine $P[X \geq 70]$ (2 points).
3. In the process of fabrication of lamps, each lamp has a probability of 0.5% of being broken. We sell lamps in boxes of 100. Let X be the random variable that counts the number of broken lamps. Determine the parameters μ and σ of this distribution (1 point).
   Then, determine the probability that in a certain box there are (0.5 point each):
   - No broken lamp.
   - Some broken lamp.
   - Exactly two broken lamps.
4. The arrival of a bus follows a probability function given by the density function

$$f(x) = \begin{cases} 0 & \text{if } x < 8 \\ x - 7.5 & \text{if } 8 \leq x < 9 \\ 0 & \text{if } x \geq 9. \end{cases}$$

   A certain person arrives at $x = 8$ to the station. What is the probability that the bus arrives before $x = 8.5$? (2 points).
   Determine the expected arrival time (1 point). What can you say about the standard deviation?

1. Let $Z \sim N(0,1)$. Then,

$$P(X \leq 3) = P\Big(Z \leq \frac{3-6}{4}\Big) = P(Z \leq -0.75) = 0.2266.$$

$$P(X \geq 12) = P\Big(Z \geq \frac{12-6}{4}\Big) = P(Z \geq 1.5) = 1 - P(Z \leq 1.5) = 0.0668.$$

$$P(5 \leq X \leq 8) = P(-0.25 \leq Z \leq 0.5) = P(Z \leq 0.5) + P(Z \leq 0.25) - 1 = 0.2902.$$

2. We can approximate this binomial distribution by

$$X' \sim N(60, \sqrt{200 \cdot 0.3 \cdot 0.7}) = N(60, 6.48).$$

Since we have a continuous model and our target was discrete, the best we can do is guessing $P(X' \geq 69.5)$ and this is just

$$P\Big(Z \geq \frac{69.5-60}{6.48}\Big) = P(Z \geq 1.47) = 0.0808.$$

3. Since each lamp has a probability of being broken of 0.5%, we have a binomial distribution for which $n = 100$ and $p = 0.005$. Then, the expectation is 0.5 and $\sigma = \sqrt{0.5 \cdot 0.005 \cdot 0.995} = 0.705$.
The probability of having 0 broken lamps can be computed just by considering $0.995^{100} = 0.606$. If we approximate the distribution by $X \sim N(0.5, 0.71)$, we will get that

$$P(X \leq 0.5) = P(Z \leq 0) = 0.5.$$

For the second item, we just want the complementary event, so the probability is again 0.5 (the "real" probability is 0.394). For the case of two broken lamps, approximating with the normal we get

$$P(1.5 \leq X \leq 2.5) = P\Big(\frac{1.5-0.5}{0.71} \leq Z \leq \frac{2.5-0.5}{0.71}\Big)$$
$$= P(1.41 \leq Z \leq 2.82) = 0.0769,$$

while the real value is $\binom{100}{2}0.995^{98} \cdot 0.005^2 = 0.0757$.

4. First of all, we need to check that this a probability distribution. We observe that between 8 and 9, probability is a straight line going from $(8, 0.5)$ and $(9, 1.5)$. Hence, we must find the area under the curve, that is identified with a trapezoid, and hence we obtain $1 \cdot \left(\frac{0.5+1.5}{2}\right) = 1$. Further, $f(x) \geq 0$ for all x.
For the probability of being between 8 and 8.5, we have another trapezoid, since now we are concerned with a line joining the points $(8, 0.5)$ and $(8.5, 1)$. Hence, we obtain $0.5 \cdot \left(\frac{0.5+1}{2}\right) = 0.375$.
The last part is a "not easy" question in which I wanted you to think. One possible approach will be trying to think in which point, say $8 + a$, we have that the probability of arriving before a is 0.5 (this corresponds to the median). Hence, we want that $a \cdot \left(\frac{1+a}{2}\right) = 0.5$. This is equivalent to $a^2 + a - 1 = 0$ and hence $a = \frac{-1+\sqrt{5}}{2} \simeq 0.61$. However, if what we want is the expectation, we must compute

$$\mu = \int_8^9 x(x - 7.5)\, dx = 8.58.$$

The variance is more subtle, and it requires to do the integral

$$\int_8^9 (x-\mu)^2(x-7.5)\,dx.$$

SEVENTH HAND IN: INTERVAL ESTIMATION

24th April 2017

This seventh hand-in is the first one devoted to statistical inference, focusing on the results concerning confidence intervals.

1. Recall the random variable of "Homework 3", that we call X and that takes 3 different values, $1, 2, 3$, with the following probability function

$$\Pr(X = 1) = 3k - 2, \quad \Pr(X = 2) = 3 - 4k, \quad \Pr(X = 3) = k,$$

where k is a certain real number betwen $2/3$ and $3/4$. We observe 100 realizations of the variable and we obtain the following frequencies:

Value of X	**Frequency**
1	10
2	15
3	75

Estimate the value of k (2 points).

2. Packets of sugar manufactured by a certain machine satisfy $\mu = 500$ g and $\sigma = 35$ g. Packets are grouped in boxes of 100 units.
Determine the probability that the mean of the weights of the packets of a box is smaller than 495 g (1 point).
Find the confidence interval of $\bar{x}$ for a probability of 95% (1 point).
Determine the probability that a box of 100 packets weighs more than 51 kg (1 point).

3. We want to estimate the mean of the results that students of a given country would obtain when solving a certain test. For that, we give the test to 400 students, and we obtain the following results.

Grade	**Frequency**
1	24
2	80
3	132
4	101
5	63

Estimate, with a confidence level of the 95% the value of the mean of the population (2.5 points).

4. The duration of a process satisfies $\sigma = 0.5$ s. What is the number of measures we must take if we want, with a confidence level of 99%, that the error is smaller than 0.1? (2.5 points).

1. We begin by computing $\bar{X}$ using the frequency table, and we obtain 2.65. Since $\mathbb{E}(X) = 4 - 2k$ (seen in “Homework 3”), we can use the moment method and write
$$4 - 2k = 2.65$$
and we obtain $k = 0.675$.

2. The mean follows a normal distribution $X \sim N\Big(500, \frac{35}{\sqrt{100}}\Big) = N(500, 3.5)$. For the first item, we want
$$P(X < 495) = P(Z < -5/3.5) = 0.0764,$$
where we have written Z for the standard normal $N(0,1)$. For the second one, the given interval will be (taken into account that $z_{0.975} = 1.96$)
$$\Big(500 - 1.96 \cdot \frac{35}{\sqrt{100}}, 500 + 1.96 \cdot \frac{35}{\sqrt{100}}\Big) = (493.14, 506.86).$$
For the last question, we have to find the probability that the mean is greater than 510 g, and with the previous notations,
$$P(X > 510) = P(Z > 10/3.5) = 0.0021.$$

3. We can use Excel to get estimates for both the mean and the standard deviation and directly get $\bar{x} = 3.622$ and $s = 1.244$. Observe that although the variance is unknown, since n is big enough we can use the Z-test. Hence, we get, since $z_{1-0.05/2} = 1.96$, that the desired interval is
$$\Big(3.622 - 1.96 \cdot \frac{1.244}{\sqrt{400}}, 3.622 + 1.96 \cdot \frac{1.244}{\sqrt{400}}\Big) = (3.500, 3.744).$$

4. We just observe that, if write E for the error,
$$E = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$
Then, in this case
$$n = \Big(\frac{z_{1-\alpha/2} \cdot \sigma}{E}\Big)^2 = \Big(\frac{2.58 \cdot 0.5}{0.1}\Big)^2 = 166.41.$$
Hence, we must perform at least 167 measures.

EIGHTH HAND IN: HYPOTHESIS TESTING

25th April 2017

In this eighth homework, we review some exercices concerning interval estimation and we start the study of hypothesis testing, via some easy examples and questions.

1. We have taken a sample of 300 people in a city (all older than 15) and we have seen that 104 regularly read the newspaper. Determine, with a confidence level of 90%, an interval to estime the proportion of readers of the newpaper among people older than 15 (2.5 points).

2. Explain in a clear way what is an error of type I and an error of type II. Which factors do they depend on? (2 points).

3. In last year's election, the 53% of the neighbors supported the president. Today, we have interviewed 360 people and 176 of them have supported the president. Can we affirm, with a confidence level of 90% that the president does not lose popularity? (2.5 points).

4. To know the failure temperature of a certain electronic component that must be installed on, we put a sensor of vibrations along the asphalt that provided the following sample of failure temperatures:

   $$77.0, 77.5, 78.0, 79.2, 79.5, 80.2, 80.7, 81.0,$$

   $$81.7, 82.1, 82.2, 82.7, 83.5, 83.3, 83.4, 84.0.$$

   From prior study it is known that the standard deviation is $\sigma = 5^{o}$.
   Determine a 90% confidence interval for average of failure temperature (1.5 points).
   Test, with a signification level of $\alpha = 0.05$ if can be accepted that the average of failure temperature is 80^{o} (1.5 points).

1. First of all, we can estimate with our data the value of the proportion of readers, say $\hat{p} = \frac{104}{300} = 0.347$. Then, by the results we have seen, and taken into account that $z_{1-0.1/2} = z_{0.95} = 1.645$, we have that the requested interval is

$$\left(\hat{p} - z_{1-\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{1-\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) = (0.301, 0.392).$$

2. There are two types of errors: a **Type I error** occurs when H_0 is rejected being true; a **Type II error** occurs when H_0 is not rejected being false. Recall that the probability of a Type I error is denoted by α (and is often called "significance level").

3. Here, H_0 represents the hypothesis that the president does not lose popularity (i.e. that the proportion of people that support him is ≥ 0.53). Hence, we have to compute a region in which the proportion of supports will lie with 90% of probability if H_0 were true. However, there is a difference between this exercice and the first one: here, if we are in the right part of the normal distribution we always accept H_0, so we want to compute that value r such that $P(X < r) = 0.1$, being $X \sim N\left(0.53, \frac{\sqrt{0.53 \cdot 0.47}}{\sqrt{360}}\right) = N(0.53, 0.0083)$. This is precisely

$$\frac{r - 0.53}{0.0083} = z_{0.1} = -1.28,$$

or what is the same $r = 0.519$. Since $\frac{176}{360} < 0.519$, we must reject H_0, or said in another way, the value $176/360 = 0.489$ is not in the acceptance region, that is $(0.519, 1)$.

4. The mean of the 16 values is 81. Since $\alpha = 0.1$, we get that $z_{1-\alpha/2} = 1.645$. Hence, the interval is

$$\left(81 - 1.645 \cdot \frac{5}{4}, 81 + 1.645 \cdot \frac{5}{4}\right) = (78.94, 83.06).$$

Observe that if σ were not given, we would not have used the Z-distribution, but a t-distribution of $16 - 1 = 15$ degrees of freedom.
If we want to test that the failure temperature is 80^{o} with a signification level of $\alpha = 0.05$, we can try to see if 80 belongs to the same interval as before but now under the assumption that $z_{1-\alpha/2} = 1.96$. The interval now is $(78.55, 83.45)$ and then the hypothesis can be accepted at that signficance level.

NINTH HAND IN: MORE ON HYPOTHESIS TESTING

26th April 2017

This last (and easy) homework assignment deals with further aspects on hypothesis testing, and must serve as a way of checking your understanding in these last topics.

1. We have flipped a coin 100 times and we have obtained 62 heads. Determine the probability of head with confidence intervals of 90%, 95% and 99% (1.5 points). Now, we want to estimate the probability of head with an error smaller than 0.002 and a confidence level of 95%. How many times shall we flip the coin? (1.5 points).

2. An airline claims that, on average, 5% of its flights are delayed each day. On a given day, of 500 flights, 50 are delayed. Test the hypothesis that the average proportion of delayed flights is 5% at the 0.01 level (3 points).

3. At one point a regional government has two policy options: cut public spending (CPS) or raise taxes (RT). Before making any decision we do a population a survey, from which the following results have been obtained:

Affiliation	**CPS**	**RT**	**Total**
A	62	90	152
B	103	85	188
C	31	29	60
Total	196	204	400

   Can we conclude at 10% of significance level that there exists no relationship between the affiliation policty and the support of the electorate to each one of the economic choices? (4 points).

1. One options consists on computing the intervals using

$$\Big(\hat{p} - z_{1-\alpha/2} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{1-\alpha/2} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\Big).$$

   Taken, into account that $p = 0.62$, we have:

   - When $\alpha = 0.1$, we get $(0.540, 0.700)$.
   - When $\alpha = 0.05$, we get $(0.525, 0.715)$.
   - When $\alpha = 0.01$, we get $(0.495, 0.745)$.

   The error is given by $z_{1-\alpha/2} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$. In this case, it will be $0.002 = 1.96 \cdot \frac{0.485}{\sqrt{n}}$ and so $n = 226270$.

2. We will use here the χ^2-test. If our hypothesis were true, the expected number of delayed flights would be $0.05 \cdot 500 = 25$, while the number of non-delayed flights would be 475. Hence, the chi valued is given by

$$\frac{(50-25)^2}{25} + \frac{(450-475)^2}{475} = 26.32.$$

   Since $\chi^2_{0.01,1} = 6.63 < 26.32$, we conclude that H_0 must be rejected.

3. If no relationship exists, the proportion of people preferring CPS would be $196/400 = 0.49$ and those preferring RT would be 0.51. Hence, the table would be as follows:

Affiliation	**CPS**	**RT**	**Total**
A	$152 \cdot 0.49$	$152 \cdot 0.51$	152
B	$188 \cdot 0.49$	$188 \cdot 0.51$	188
C	$60 \cdot 0.49$	$60 \cdot 0.51$	60
Total	196	204	400

Computing all the numbers, we obtain:

Affiliation	**CPS**	**RT**	**Total**
A	74.48	77.52	152
B	92.12	95.88	188
C	29.4	30.6	60
Total	196	204	400

Hence, the chi value will be

$$\frac{(62-74.48)^2}{74.48} + \frac{(90-77.52)^2}{77.52} + \frac{(103-92.12)^2}{92.12}$$
$$+\frac{(85-95.88)^2}{95.88} + \frac{(31-29.4)^2}{29.4} + \frac{(29-30.6)^2}{30.6}$$
$$= 6.77$$

We see that $\chi^2_{0.1,6-1} = 9.24$, and since our value was smaller, we accept H_0.