PL 2

Probabilities and Random Variables

2.1 Conditional probability, independence

Answer the following questions through simulations in Matlab and whenever asked compare the results obtained with the theoretical values:

1. Consider families with children where the probability of boys being born is equal to that of girls being born:

   (a) Obtain a simulation estimate of the probability of the event "having at least one boy child" in families with 2 children.

   (b) Determine the theoretical value of the previous event and compare it with the estimate obtained by simulation. Are the values the same? Why?

   (c) Suppose that for a family with 2 children chosen at random, we know that one of the children is a boy. What is the probability of the other child being also a boy? Determine the theoretical value of this probability and estimate the same probability by simulation.

   (d) Knowing that the first child is a boy in a family with 2 children, determine by simulation the probability of the second child being also a boy. What can be concluded from the obtained result regarding the independence of events?

   (e) Consider a family with 5 children. Knowing that at least one of the children is a boy, obtain by simulation an estimate for the probability that one of the others (and only one) is also a boy.

   (f) Repeat question (e), but estimating the probability that at least one of the others is also a boy.

2. Consider the following "game": blindfolded launch of n darts, one at a time, to m targets, ensuring that each dart always hits a target (and only 1).

   (a) Estimate through simulation the probability that no target is hit more than once when $n = 20$ darts and $m = 100$ targets.

   (b) Estimate through simulation the probability that at least 1 target is hit 2 or more times when $n = 20$ darts and $m = 100$ targets.

   (c) Consider the values of $m = 1000$ and $m = 100000$ targets. For each of these values, make the necessary simulations to draw a graph (using the Matlab *plot* function) of the probability of question (b) as a function of the number of darts n. Consider n from 10 to 100 in increments of 10. The 2 graphs must be sub-graphs of one same figure (use Matlab's *subplot* instruction). Compare the results of the 2 cases and draw conclusions.

   (d) Consider the value of $n = 100$ darts. Make the necessary simulations to draw a graph of the probability of question (b) as a function of the values of m = 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000 targets. What do you conclude from the obtained results?

3. Consider an array of size T that serves as the basis for the implementation of an associative memory (for example in Java). Assume that the *hash* function returns a value between 0 and $T - 1$ with all values equally probable.

(a) Determine by simulation the probability of having at least one collision (at least 2 *keys* mapped by the *hash* function to the same position in the array) if 10 *keys* are inserted into an array of size $T = 1000$.

(b) Draw a graph with the probability of question (a) (estimated by simulation) as a function of the number of *keys* for all relevant values in an array of size $T = 1000$.

(c) For a number of 50 *keys*, draw a graph with the probability (estimated by simulation) of not having any collisions as a function of the array size T (assume T sizes from 100 to 1000 with increments of 100).

4. Consider a party where a given number n of persons is present.

   (a) What should be the lowest value of n for which the probability of two or more persons having the same birthday (month and day) is greater than 0.5 (assume that a year always has 365 days)?

   (b) What should be the value of n for the probability of the previous question to be greater than 0.9?

5. Consider a six-sided die (numbered from 1 to 6) thrown 2 times. Assume that the die is balanced (same probability for all faces to be up). Consider the following events: “A - the sum of the two values is equal to 9”, “B - the second value is even”, “C - at least one of the values is equal to 5” and “D - none of the values is equal to 1”.

   (a) Estimate by simulation the probability of each of the 4 events.

   (b) Determine theoretically if events A and B are independent.

   (c) Determine theoretically if events C and D are independent.

6. Consider a language with only 3 words {“one”,“two´´,“three”} and which allows sequences of 2 words. Consider that all sentences have the same probability and that the two words in a sentence can be the same. The answers to the following questions should be based on theoretical values.

   (a) What is the probability of the sequence “one two”?

   (b) What is the probability of “one” to appear at least once in a sequence?

   (c) What is the probability of a sequence with “one” or “two”?

   (d) What is the value of P[“sequence includes the word one”| “sequence includes the word two”]?

7. Consider a company with 3 programmers (André, Bruno and Carlos). Assume that the probability of a program of each of them having “bugs” and the number of programs developed by each of them are the values shown in the following table.

Programmer	Prob(“bug”)	No. of programs
André	0.01	20
Bruno	0.05	30
Carlos	0.001	50

   The Director of the company randomly selects one of the 100 programs developed by its 3 programmers and discovers that it contains a bug.

   (a) What is the probability that the program is from Carlos?

   (b) Whose programmer is most likely to be the program selected by the Director?

2.2 Random variables and distributions

1. Consider the random variable X corresponding to the face up on the roll of one die. Using theoretical values:

   (a) produce a graph, in Matlab, that represents the mass probability function [1] of X;

   (b) on a second graph of the same figure, draw the graph of the cumulative probability distribution function [2] (use function `stairs` of Matlab).

2. Consider a box containing 90 bills of 5 Euros, 9 bills of 50 Euros and 1 bill of 100 Euros.

   (a) Describe the sampling space of the random experiment, ”remove a bill from the box”, and the probabilities of the elementary events.

   (b) Now consider the random variable X as the value of a bill randomly taken from the box described above. Describe the sampling space of X and its probability mass function.

   (c) Determine the cumulative probability distribution function of X and show its graph in Matlab.

3. Consider that a fair coin is flipped 4 times. Let X be the random variable representing the number of tails observed in the 4 flips.

   (a) Simulate the experiment and estimate the mass probability function $p_X(x)$ of the random variable X.

   (b) Based on $p_X(x)$, estimate the expected value, variance and standard deviation of X.

   (c) Identify the type of distribution of the random variable X and write the theoretical expression of its probability function.

   (d) Calculate the theoretical values of the probability mass function of X and compare them with the estimated values obtained, using simulation, in (a).

   (e) Calculate the theoretical values of $E[x]$ and $Var(X)$ and compare them with the values obtained in (b).

   (f) Based on the theoretical values of the probability mass function of this distribution, calculate:

      i. the probability of obtaining at least 2 tails;

      ii. the probability of obtaining up to 1 tail;

      iii. the probability of obtaining between 1 and 3 tails.

4. Knowing that a manufacturing process produces 30 % of defective parts and considering the random variable X, which represents the number of defective parts in a sample of 5 parts taken at random, obtain:

   (a) By simulation:

      i. an estimate for the mass probability function of X;

      ii. the graph representing the cumulative probability distribution function of X;

      iii. estimate for the probability that, at most, 2 of the parts of a sample are defective.

   (b) Analytically:

      i. the cumulative probability distribution function of X;

      ii. the probability that, at most, 2 of parts in a sample are defective.

[1] The mass probability function is often referred to simply as the probability function

[2] The cumulative probability distribution function is often referred to simply as the distribution function

5. Assume that the engine(s) of an airplane can fail with probability p and that failures are independent between engines. Assume also that the plane crashes if more than half of the engines fail. Under these conditions, do you prefer to fly on a plane with 2 or 4 engines? Use the distribution that you consider the most appropriate.

   Suggestion: You have at least 2 alternatives: (1) obtain expressions for the probability of each type of plane crashing as a function of p and use the quotient between both to answer the question, (2) perform the calculations to a set of specific values[3] of p (ex: `p = logspace(-3,log10(1/2),100)`) and use a graph showing simultaneously the crashing probabilities of each type of plane.

6. The Poisson distribution is a limit form of the binomial distribution (when $n \to \infty$, $p \to 0$ and np remains constant) and therefore can be used to approximate and simplify the calculations for the binomial distribution in situations where the previous conditions are meet.

   In an industrial chip manufacturing process, some of the chips are defective making them unsuitable for commercialization. It is known that, on average, there is one defective chip for every 1000 chips.

   (a) Using the binomial distribution, determine the probability that a sample of 8000 chips has 7 defective ones.

   (b) Determine the same probability using the Poisson approximation and compare the result with the previous one.
   Poisson mass probability function: $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$

7. Assume that the number of messages arriving at an *email* server follows a Poisson's law with an average of 15 (messages per second). Calculate the probability that within one second:

   (a) the server does not receive any messages;

   (b) the server receives more than 10 messages.

8. Assuming that the number of typographical errors on a book page has a Poisson distribution with $\lambda = 0.02$, calculate the probability that there is a maximum of 1 error in a 100 page book. Consider that the number of errors on each page is independent of the number of errors on the other pages.

9. Considering the random variable X, which represents the student grades in a given course, continuous [4] and with normal distribution [5] (mean 14 and standard deviation 2), obtain through simulation an estimate for the probabilities of:

   (a) a student in the course having a grade between 12 and 16;

   (b) students having grades between 10 and 18;

   (c) a student passes (grade greater than or equal to 10);

   (d) check the correctness of the previous results using the Matlab `normcdf()` function.

[3] Run `help logspace` in Matlab to understand the `logspace` arguments used in the example.

[4] Equivalent to considering that the grades are real numbers.

[5] Use the Matlab function `randn()`.

2.3 Section for evaluation [6]

Consider a toy manufacturing company producing a given toy. The toy is composed by two components (1 and 2) that are produced separately and then are assembled together. At the end, the toys are packed for commercialization in boxes of n toys each.
The manufacturing process of Component 1 produces $p_1 = 0.2\%$ of defective components. The manufacturing process of Component 2 produces $p_2 = 0.5\%$ of defective components. A toy is defective if at least one of its components is defective. The assembly process produces $p_a = 1\%$ of defective toys (even when none of the 2 components is defective).

1. **(Evaluation weight = 20%)** Consider the event "A – a box of toys has at least 1 defective toy".

   (a) Estimate by simulation the probability of event A when $n = 8$ toys.

   (b) Estimate by simulation the average number of toys that are defective only due to the assembly process when event A occurs.

2. **(Evaluation weight = 30%)** Consider the event "B – a box of toys has no defective toys".

   (a) Estimate by simulation the probability of event B when $n = 8$ toys. Check the consistency of this result with the one obtained in question 1(a).

   (b) Determine the theoretical value of the probability of event B and compare it with the value estimated by simulation in question 2(a). What do you conclude?

   (c) Make the necessary simulations to draw a *plot* graph of the probability of event B as a function of the box capacity n. Consider all values of n from 2 to 20. Describe and justify the obtained results.

   (d) Analysing the plot drawn in the previous question 2(c), what must be the maximum box capacity if the company wants to guarantee that the probability of each box having no defective toys is at least 90%?

3. **(Evaluation weight = 30%)** Consider the random variable X representing the number of defective toys in a box.

   (a) Estimate by simulation the mass probability function $p_X(x)$ of X when $n = 8$ toys and draw it in a *stem* graph. Describe the obtained results and check their consistency with the result obtained in question 2(a).

   (b) Based on $p_X(x)$, compute the probability of $X >= 2$. What do you conclude?

   (c) Based on $p_X(x)$, estimate the expected value, variance and standard deviation of X.

   (d) Repeat questions 3(a), 3(b) and 3(c) but now considering $n = 16$ toys. Compare all results with the previous ones (obtained with $n = 8$ toys) and justify the differences.

4. **(Evaluation weight = 20%)** Assume now that the company aims to commercialize the toys in boxes of $n = 20$ toys guaranteeing that the probability of a box without defective toys is at least 90%.

   To reach this goal, the assembly process was improved by reducing p_a to 0.1% and a quality assurance process was implemented as follow: a sample of m toys (with $1 \leq m < 20$) is selected from each box for testing; the box is not commercialized if at least one of the selected toys is defective or is commercialized, otherwise.

   (a) Estimate by simulation the probability of a box being commercialized when the quality assurance process is set with $m = 1$ (check the usefulness of Matlab function $randperm$ in the implementation of the simulation). What do you conclude?

   (b) Estimate by simulation the lowest value of m that is required to reach the desired goal.

[6] The execution of this section is for evaluation. You should make a report (to be submitted in PDF) as complete as possible with the answers to all questions in this section. The report should start by identifying the academic year, the course unit, the practical class and the elements of the group (name and student number) that carried out the work. Whenever you need to implement a Matlab code, you must include the code in the report properly commented.