Modelação de Sistemas Complexos
Departamento de Física
Universidade de Aveiro

Project 1

Probability. Binomial and Poisson distributions. The central limiting theorem.

Nuno Monteiro (nº mec 79907)

12-03-2020

Probability

Task 1.1

A random number x can take the values 1,2,3, ... M uniformly at random with the same probability 1/M. Prove analytically that the averaged value <x> of the random number x is

$$< x >= \frac{M+1}{2} \tag{1}$$

and the variance is

$$< (x- < x >)^2 >= \frac{M^2-1}{12} \tag{2}$$

To calculate anallyticaly the averaged value of x, we first recognize that $x \in [1, M]$. And if each possible number inside that interval has the same probability of 1/M (uniform distribution), then to get <x>, it's necessary to sum all the values inside [1,M] (with an integral) and devide by the length of the interval.

$$< x >= \int_1^M \frac{x}{M-1} dx = \frac{1}{M-1} \left[\frac{x^2}{2}\right]_1^M = \frac{M^2-1}{2(M-1)} = \frac{M+1}{2}$$

Now, for the analytical approach of the variance of x, firstly, one should start with the simplification of $< (x- < x >)^2 >$:

$$\begin{aligned} \sigma^2 &=< (x- < x >)^2 > \\ &= \frac{1}{M} \sum_i (x_i^2 - 2 < x > x_i + < x >^2) \\ &= \frac{\sum_i xi^2}{M} - \frac{2 < x > \sum_i xi}{M} + \frac{M < x >^2}{M} \\ &= \frac{\sum_i xi^2}{M} + 2 < x >^2 + < x >^2 \\ &=< x^2 > - < x >^2 \end{aligned}$$

Then, by using the following mathematical property

$$\sum_{i=1}^{n} l_i^2 = \frac{1}{6} n(n+1)(2n+1) \tag{3}$$

we can replace $< x^2 >$ with (3)

$$< x^2 >= \frac{1}{M} \sum_{i=1}^{M} x_i^2 = \frac{1}{M} \frac{1}{6} M(M+1)(2M+1) = \frac{1}{6}(M+1)(2M+1) =$$

$$= \frac{1}{6}(2M^2 + 3M + 1)$$

Hence, applying the first proof, $< x >= \frac{M+1}{2}$, we get

$$var(x) = \frac{1}{6}(2M^2 + 3M + 1) - \frac{M^2 + 2M + 1}{4} = \frac{1}{12}((4-3)M^2 + 0M + 2 - 1) = \frac{1}{12}(M^2 - 1)$$

As we wanted to prove.

Prove analytically that the probability that x is smaller or equal q=60 is 0.6 at M=100.

In probability and statistics, the quantile function, associated with a probability distribution of a random variable, specifies the value of the random variable such that the probability of the variable being less than or equal to that value equals the given probability. So, interpreting the last paragraph, q is by definition a quantile.

To find the probability, knowing how the set of random variables behaves, by the cumulative probability, one can find the q^{th} quantile of a random population. So, in this case as we have a uniform distribution and values range between 0 and 100, the cumulative distribution is:

$$F(x) = \frac{x}{100 - 0} = \frac{x}{100}, 0 \leq x \leq 100 \tag{4}$$

Simply replacing $x = q = 60$ in $F(x)$, the probability, hence, is $p = 60/100 = 0.6$, as we wanted to prove.

Task 1.2

In this task we prove eqs. (1) and (2) by applying the algoritmn given in the proposal for this project. Basically the algorithm consists in finding the average value and the variance, numerically. Then, to see the law of large numbers in practice, we compare the numerical reaults, raising the number of attempts ($N = 1000, 10^4, 10^6$), with the anallitical values (Task 1.1), which are theoretical.

For this computation, the chosen M is M=100. Hence, the exact values of the average and variance are, respectivelly, 50.5 and 833.25 (Eqs. (1) and (2)). The results for a random seed of N attempts each time are displayed bellow. To check the accuracy in respect to the exact values, the relative errors are performed in the last two lines of the following table:

N	10^3	10^4	10^6
$< x >$	51.0130	50.8046	50.5016
var(x)	859.4568	821.6430	832.5415
RE $< x >$ (%)	1.0158	0.6032	0.0033
RE var(x) (%)	3.1451	1.3930	0.0850

Table 1: Means and Variances in function of N attempts of generated values obeying to a uniform probability distribution. Last 2 lines are the respective relative errors, given by $(|x_{exact} - x_{measured}|)/x_{exact} \times 100$ (%)

The law of large numbers, which states that the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed. As we see in table 1, the results from the sets of random numbers of the mean and variance converge to the correct values as N increases, and this law is comproved here. Also, the average value $< x >$ shows better accuracy because all the relative errors (%) are lower, whereas variance's accuracy is worse in an order of degree. This can be explained by the nature of the calculations performed, as the mean value obey to a linear equation and the variance to a parabolic equation, which is more prone to bigger variations.

Task 1.3

Prove that if two random numbers x= rand(1) and y= rand(1) are uncorrelated random variables then the mean value of a random variable z=xy is <z>=<x><y>.

To prove this, the means of x, y and z are computed, and then the calculation of <x><y> is compared to <z> (mean of z). The number of attempts performed are $N = 1000$, 10^4, 10^6. For the computed summations, x_i and y_i are both randomly generated and can take, with a unif. distribution, values from 0 to 1, uniformly, while $z_i = x_i y_i$. Even if the two random sets (x and y) of random variables are uncorrelated, if they have a uniform probability and range from 0 to 1 (as z does), from the law of large numbers (stated in task 1.2), as N progresses, the calculation $< x >< y >$ should tend to $< z >$.

The table bellow shows the results for varying N, of the he exact calculation of $< z >$ and the proposed calculation of the mean ($< z >=< x >< y >$). The third line corresponds to the accuracy given by the relative error. We see that the accuracy improves as N is higher, and is always lower than 10%, we prove here the previous statement, especially for larger numbers.

N	10^3	10^4	10^6
$< x >< y >$	0.2533	0.2506	0.2498
mean($< z >$)	0.2614	0.2509	0.2497
Acc. (%)	3.1000	0.1041	0.0345

Table 2: Results of the average values and variances of z varying N.

Probability density

Task 2.1

Find numerically the probability density p(x) for a random variable x=rand(1), i.e. x is a random number in the interval $0 \leq x < 1$ *. Plot* $p(x)$*.*

By generating, in a software program, 10,000 random numbers rand(1), the probability of each one is then calculated. As these numbers all have the same probability, the result is theoretically a line with no slope (m=0). In figure 2 (b), the probability density of the discrete values approximatelly represents the previously described function, with values oscillating around p(x)=0.03, for an interval dx of 1/30 in the dataset. This result makes sense, as the number of elements inside each bin would have, in practice, the same probability of occurring.

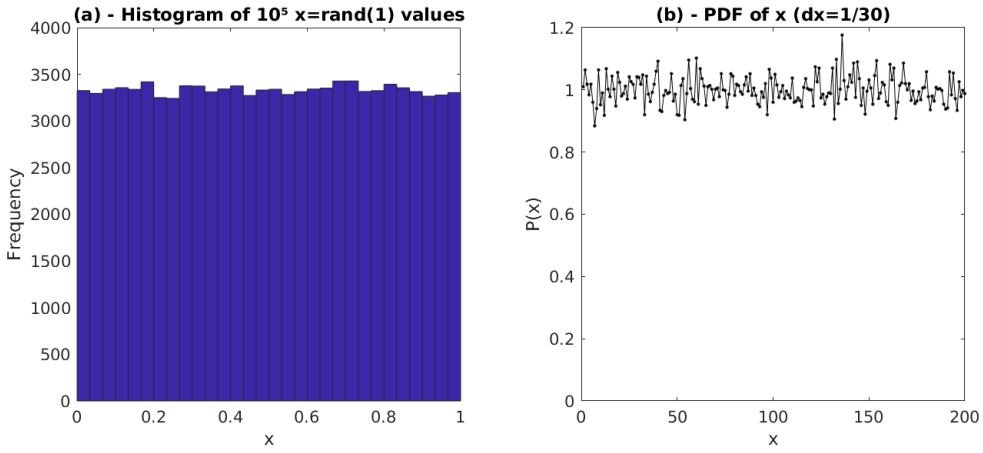

Figure 1: (a) Histogram, and (b) Probability distribution function of rand(1) generated values

Task 2.2

Find the mean value $< x >$*, and the variance* $\sigma^2 = \int_0^1 (x - < x >)^2 p(x) dx$*. Compare with the theoretical values* $< x >= 1/2$*,* $\sigma^2 = 1/12$*.*

The results for a given computational random seed for the 100,000 generated random numbers are:

- $< x >= 0.5011$, with an accuracy* of 0.2175%
- $\sigma^2 = 0.0828$, with an accuracy of 0.6085%

(*measured by the relative error)

These results are satisfying, and confirms that the algorithm implemented for the generation of x uniformly distributed numbers was correct.

Task 2.3

Find the probability density g(x) of a random variable x defined as $= \sqrt{rand(1)}$. *Plot* $g(x)$ *versus* x. *Compare with the theoretical result:* $g(x) = 2x$.

The algorithm used for this task consists in simply generating different probabilities inside the width of each bin ($\Delta x = 0.005$), inside the interval of possible values [0,1]. The length of each set of numbers varies also varies $N = 100, 10^3, 10^6$ iterations, so it is possible to see the effects of more iterations of random numbers.

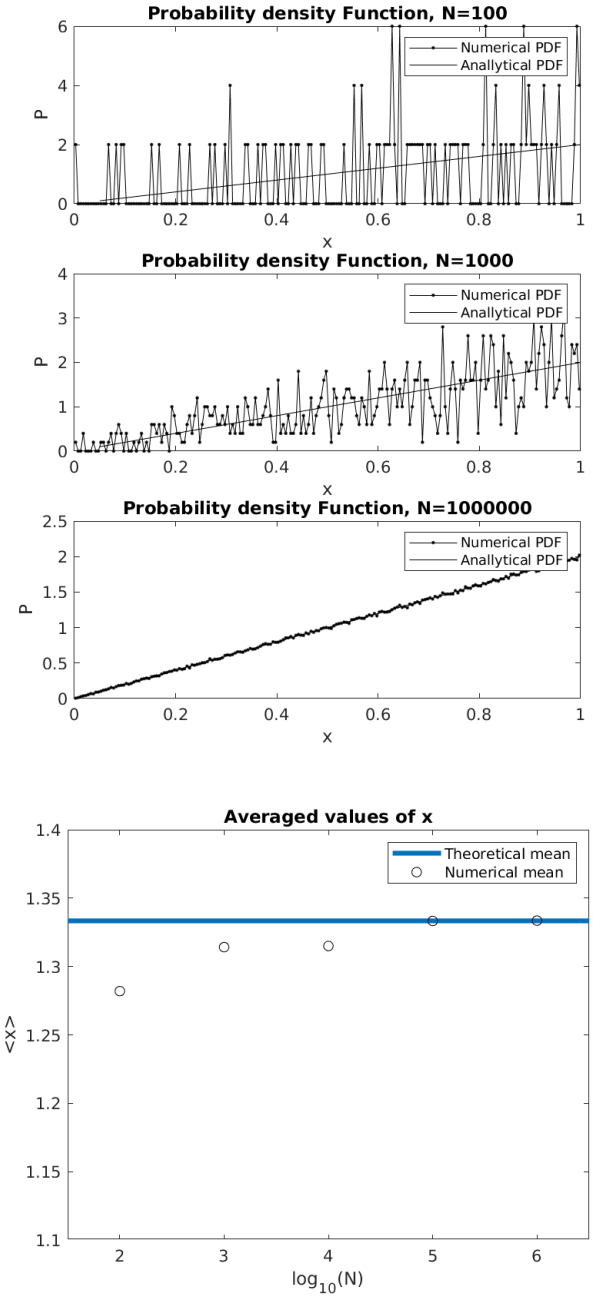

Figure 2: The 3 upper plots show the probability density functions, in function of the center of each bin of the data set (for N=100, 10^3, 10^6). The last plot represents the averaged values of x=$\sqrt{rand(1)}$ (for N=100, 10^3, 10^4, 10^5, 10^6), and the theoretical prediction

Firstly, the normalization condition (the sum of all probabilities must be 1) is checked. In figure 2, the plot of the density function occurs as expected, and, as N grows, the resulting points of the probability of random numbers being in each bin converges to $g(x) = 2x$. The mean values also show this trend, as the averages in figure 2's last plot tend exponentially to the exact prediction, given by

$$< x_e >= \frac{1}{K} \sum_{k=0}^{k_{max}} g(x_k) = 1.3334$$

with $x_k = \sqrt{(k + 0.5)\Delta x}$. The accuracy for $N = 10^6$ is very good with the relative error being 0.0064%.

The central limiting theorem

Generate n random numbers z i with the mean value <z> and the variance σ^2. For example, you can use the random number generator z = rand(1). A random number Y is defined as a sum of random numbers z_i ,

$$Y = \frac{1}{n}\sum_{i=1}^{n} zi \tag{5}$$

Find the distribution function P (Y) of the random numbers Y. Plot P (Y) . Shows that the mean value of Y is equal to <z>. Calculate the variance of Y,

$$\Lambda^2 =< (Y - < Y >)^2 > \tag{6}$$

and show that Λ^2 tends to σ^2/n when the number of trials tends to infinity.

For this exercise, an algorithm (with detailed information in the proposal sheet of this project) is given for implementation of the random numbers Y (equation 5). Following the instructions, this random set of variables, Y, obey to a normal distribution. The parameters chosen are: n=10, 100, 1000 (number of means calculated to get Y); $\Delta y = 0.005$, and the number of trials is $N = 10^6$.

This algorithm asks to calculate the probability distribution function $P(y_k)$,

$$P(y_k) = \frac{M(k)}{N\Delta y}$$

and present graphically this function for n=10, 100, 1000.

All distributions represent a "bell-shaped" normal distribution behavior, as expected, because each set of random uniformly generated numbers is averaged, as shown in figure 3. Therefore, for example, the resulting maximum probabilities, should represent the typical means, where the most frequent one (higher probability point) should be around 0.5, and that is what happens in all the curves of probability.

Finally, the results in figure 3 (b) represent the convergence of the variance squared (Λ^2) with the standard deviation squared divided by n (σ^2/n). For n=1000, the differences are almost none.

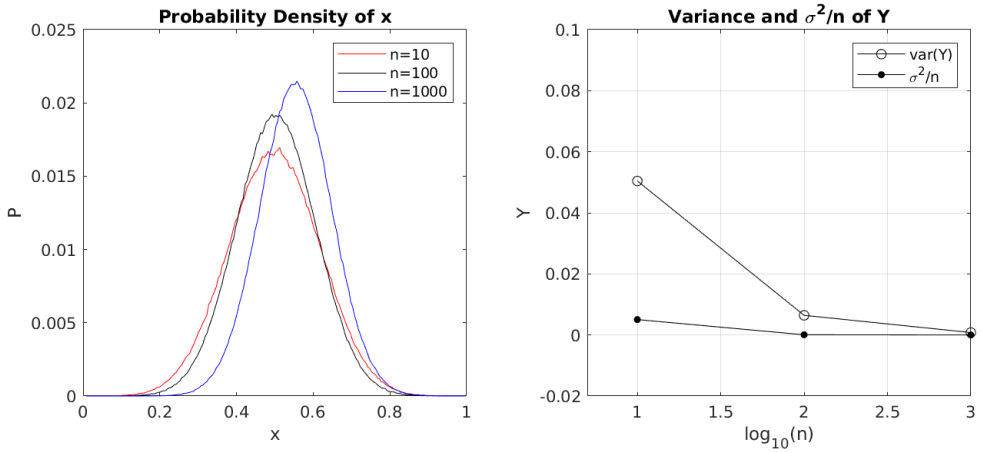

Figure 3: The left side plot shows the probability distributions and the left side plot shows the converging of Λ and σ^2/n

Throwing balls

There are M boxes and N balls. The balls are distributed among the boxes uniformly at random. Find the probability to find n balls in a given box. M=9, N=21, the number of trials K= 10^3 , 10^4 , 10^6.

An algorithm is given in the project's proposal. It asks for the generation of N (=21) x_i integer numbers each corresponding to the box where the ball landed (1 to 9), and repeat it K times. Then, it is asked to calculate the probability to find n balls in the box of number 3 as follows:

$$P(n) = \frac{N_{tr}(n)}{K} \tag{7}$$

The resulting probability distributions, varying K, given by the last equation, are given in the figure 4 (two plots), in function of the number of the number of balls.

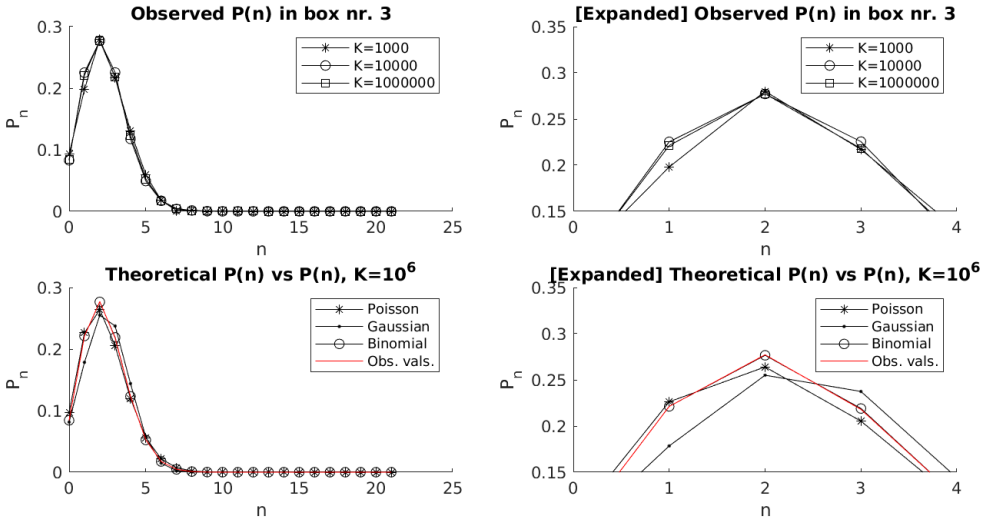

Figure 4: The upper plots show the probabilities of finding n balls in box number 3 (for K=10^3, 10^4 10^6). The last two plots represent analytical models of Poisson, Gaussian and Binomial probability distributions and the comparison with the results for K=10^6 trials.

The correlation coefficient calculation is a way to verify numerically which distribution (Poisson or Gaussian) is the best representation of the probability distribution of P(n). The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. Now we present the algorithm of the correlation coefficient (r) is the following:

1. We start by some preliminary calculations:
   - Calculate <x>, the mean of all of the first coordinates of the data x_i;
   - Calculate <y>, the mean of all of the second coordinates of the data y_i;
   - Calculate σ_x, the standard deviation of the data xi;
   - Calculate σ_y, the standard deviation of the data yi.

2. Use the formula $z_{x,i} = \frac{(x_i - <x>)}{\sigma_x}$ to get a standardized value for each x_i.

3. Use the formula $z_{y,i} = \frac{(y_i - <y>)}{\sigma_y}$ to get a standardized value for each y_i.

4. Multiply corresponding standardized values: $z_{x,i} z_{y,i}$

5. Add the products from the last step together.

6. Divide the sum from the previous step by $n - 1$, where n is the total number of points in our set of paired data. The result of all of this is the correlation coefficient r.

As seen in the previous figure, visually the best fit of the observed probability distribution is for K=10^6 trials, as expected by the law of large numbers. So, a computation of the correlation coefficient is computed, between the practical and theoretical results. The correlation coefficient between the Poisson distribution and the observed results is 0.9983,, between the Gaussian and the observations is 0.9907 and finally the correlation coefficient for the comparison with the Binomial is 1.0000, i.e. a value very near to 1. As the correlation is higher, we conclude that the Binomial distribution is the best anallytical model to represent this kind of problem, for a large number of trials of course.