Estimation of the experimental probability of the occurrence of 3 large-magnitude earthquakes on the same date of the year in Mexico

Mexico is a country where many earthquakes occur, and some of them have caused catastrophic damage and many deaths. On September 19, 1985, an earthquake of great magnitude occurred (in Mexico City) that caused terrible damage and many deaths, leaving the date marked in the memory of Mexicans. So much so, that this date was chosen to carry out an annual prevention exercise (simulation) in the event of a major earthquake.

Since then, two more large earthquakes have occurred on the same date: September 19, 2017 (Mexico City) and September 19, 2022 (Michoacán).

The objective of this exercise is to estimate the experimental probability of this coincidence and compare it with some theoretical results that are handled in various investigations. The idea is to verify informally if the theory effectively approaches what happens in reality.

Theoretical probability vs. experimental probability

Theoretical probability is based on mathematical reasoning and represents the expectation that an event will occur.

For its part, the experimental probability is based on what really happens, that is, it is calculated considering the frequency of occurrence of a given event.

Probability estimation

The calculation of the theoretical probability of this coincidence can be done in several ways depending on the considerations that are taken. We can find in the academic literature results that vary from one another according to the initial approaches of the researchers, such as the type of distribution used, time intervals, geographical restrictions, type of earthquakes, damage caused, magnitude, among many other variables. What is interesting about these studies is not so much the initial hypotheses and the different results found, but rather that they all agree that the probability of occurrence is extremely low.

To carry out both theoretical and practical calculations, it is necessary to define some initial premises in advance. In the case of this study:

- Seismic events are independent of each other.
- A uniform distribution is assumed for the earthquakes in the year.
- A major earthquake is one greater than or equal to 7 on the Richter scale. It should be noted that in the investigations carried out by seismic specialists, the term "catastrophic earthquake" or "catastrophic event" is used, where the earthquakes that have caused material damage and human losses are counted.

- Seismic events and calendar dates are independent.
- Leap years were not considered.
- The time interval used is from January 1, 1900, to September 21, 2022.

Important semantic clarifications

We can state the problem in different ways and each of them has a particular form of calculation. In this project, 2 different cases are considered:

- Case 1: "A" is the event in which a large earthquake occurs, specifically on the date September 19, three times in different years. In this scenario, the probability of occurrence of the earthquake per se is not calculated, but it is assumed that it occurs and what is verified is the possibility of it occurring on a given date.
- Case 2: "A" is an earthquake of great magnitude; therefore, P(A) is the probability that it occurs regardless of the day. In this case, the two events that are involved must be differentiated: A (earthquake occurrence) and B (September 19) on three occasions.

Theoretical calculation of probabilities

- Case 1: it is the simplest way to do the calculations. Earthquakes of great magnitude are assumed to occur, and we must live with that fact. Therefore, when they occur, we want to know the probability that 3 will occur on the same date of the year. The probability of an earthquake occurring on a particular date (e.g., September 19) is 1/365 since there are 365 days in a year and all are equally likely. Therefore, the probability P(A) that three large earthquakes occur on September 19 is:

$$P(A) = \left(\frac{1}{365}\right)^3 = \mathbf{2.1x10^{-8}}$$

- Case 2: It is not a purely theoretical calculation because it also involves real data. There are two events of interest, and we need both to occur at the same time, that is, we need the joint probability of A (occurrence of a large earthquake) and B (that the date is September 19), or what is the same P(AB). Since A and B are independent events, P(AB) = P(A)P(B). We calculate P(A) in a rudimentary way using Laplace's formula:

$$P(A) = \frac{favorable\ events}{total\ events} = \frac{87}{41{,}228} = 0.00211 = 2.11x10^{-3}$$

Where 87 is the number of earthquakes of magnitude 7 or more that have occurred since 1900 and 41,228 is the total number of earthquakes that have occurred in the same time interval (earthquakes of magnitude greater than or equal to 4. Earthquakes of lesser magnitude were ruled out because they do not cause damage according to the Richter scale). Then we calculate P(AB):

$$P(AB) = P(A)P(B) = (2.11x10^{-3})\left(\frac{1}{365}\right) = 5.8x10^{-6}$$

Finally, the probability of occurrence of three earthquakes on September 19 is obtained:

$$P(AB)^3 = (5.8x10^{-6})^3 = \mathbf{1.95x10^{-16}}$$

Although both are extremely unlikely, the difference in orders of magnitude between the two numbers is remarkable. The reason is due to the approach of each calculation: in the first, earthquakes are considered to occur, and from that fact, the probabilities are estimated. This calculation is not unlike having 365 numbered balls in a bag and drawing the same ball in three attempts. Nor does it consider the propensity for the occurrence of earthquakes in a given area; Mexico is not the same as Brazil, which is a practically aseismic area (a place where very few earthquakes occur, and those that do occur are of low magnitude). It is a purely theoretical calculation.

In the second case, the seismic nature of the region where the calculations are made is considered. It is based on the uncertainty that the earthquake(s) will occur, and therefore, this probability must be used in the calculations. Hence probability is reduced so dramatically.

It should be noted that there are several ways to estimate case 2. For example, in (Jaimes and Garcia-Soto, 2019), P(A) is replaced by an "Annual probability of return" that calculates the probability of occurrence of a catastrophic earthquake per year using PSHA (Probabilistic Seismic Hazard Analysis). If we were to use that data (p=0.094 for 2018) instead of $(\frac{87}{41{,}228})$ we would get:

$$P(AB)^3 = [P(A)P(B)]^3 = [(0.094)\left(\frac{1}{365}\right)]^3 = \mathbf{1.71x10^{-11}}$$

Let us note that although this “annual probability of return” is a much better worked data and that it considers multiple factors, the result is still very small.

Could conditional probability be used to estimate occurrence?

It is tempting to use the conditional probability formula to do the calculations, after all, we want to calculate the probability that given an earthquake, it will strike on September 19. However, conditional probability does not give us relevant information when the events involved are independent of each other, as in this case.

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

That is, as the occurrences of earthquakes do not depend on a particular day, P(A|B) = P(A). Or what is the same, we are only interested in calculating the probability of a particular date.

Calculations that resemble but do not correspond to the probability of 3 earthquakes on Sept. 19

1. $P(A) = 365(\frac{1}{365})^3 = 7.51x10^{-6}$. In this case, the probability of 3 earthquakes occurring on the same date is calculated without specifying which date. Note that it also corresponds to the probability that 2 earthquakes occur on the date September 19.

2. The birthday paradox. It asks for the probability that, in a set of n randomly chosen people, at least two will share a birthday. It could be erroneously argued that exchanging people for earthquakes and birthdays for earthquake occurrence dates would be a solution to the problem posed. The probability equation would be:

$$P(n) = \frac{365!}{365^n (365 - n)!}$$

However, this reasoning is not correct because what it would really calculate would be the probability that 2 or more earthquakes occur on September 19 considering a group of n earthquakes, that is, it is the sum of P(2 earthquakes on the same day) + P (3 earthquakes on the same day) + … + P(n-1 earthquakes on the same day). This is a big probability despite our intuition telling us otherwise.

Calculation of the experimental probability

For the calculation of the experimental probability, earthquakes of magnitude greater than or equal to 4 from January 1, 1900, to September 21, 2022, were considered. For this, the catalog of earthquakes of the National Seismological Service (SSN) of the UNAM in Mexico was taken.

The first thing to do to estimate the experimental probability is to verify how the earthquakes are distributed throughout the year, and this is done with a histogram, where the x-axis represents the days of the year, and the y-axis represents the frequency with which earthquakes occur.

Based on this distribution, the probability density function (PDF) that best approximates the data and that allows us to later calculate the probability is estimated. To estimate the PDF from the data, we use KDE (Kernel Density Estimation), which is a non-parametric method that estimates the density function of a random variable from a non-negative function called Kernel:

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \frac{(x - x_i)}{h}$$

where K is a non-negative function called Kernel, and h > 0 is a parameter called bandwidth.

In this project, all estimations were done using Python. For the estimation of the PDF, the sns.displot() function of the seaborn library was used. When the parameter kde=True is passed to this function, it uses the normal density function as Kernel by default and automatically searches for the best value of h that minimizes the mean square error, which is very convenient since it optimizes the PDF.

Figure 1 shows the histogram of all the earthquakes that have occurred in Mexico since 1900 and its PDF. Figure 2 shows the histogram of the large-magnitude earthquakes (87 in total) with their PDF. We can see that the earthquakes follow an approximately uniform distribution, which confirms that the occurrence of any of them is equally likely for every day of the year.

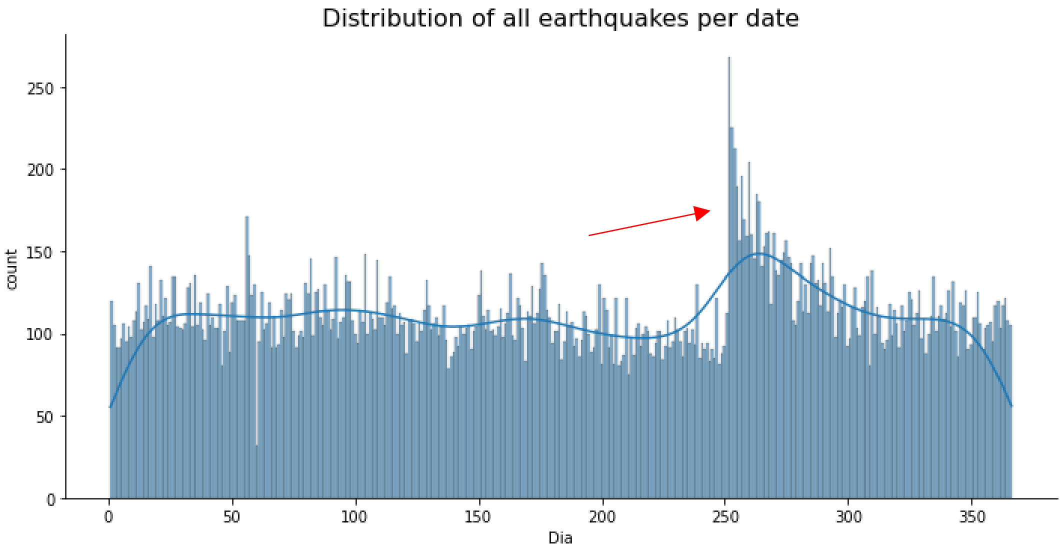

Figure 1. The distribution of earthquakes is approximately uniform with a peak in September.

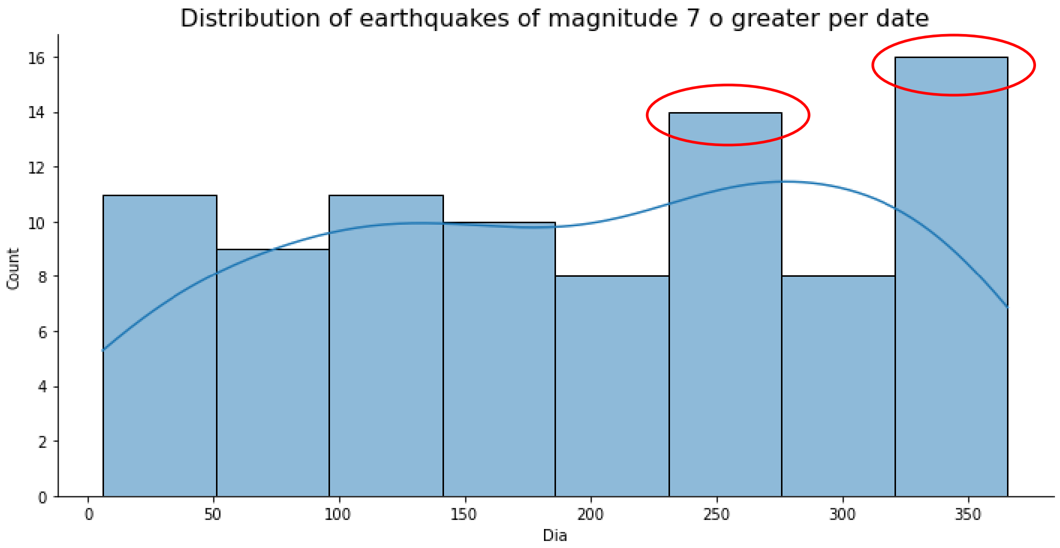

Figure 2. The distribution of large earthquakes is approximately uniform with a peak in September and another at the end of the year.

We can also note that there is a peak in the frequency of earthquakes during the month of September, which makes the PDF greater in that area, and therefore, the probability of occurrence of earthquakes during that month is slightly greater than the rest of the year.

Since the PDFs are continuous and we want to calculate the probability of a large earthquake occurring on September 19 (figure 3), we must calculate the cumulative probability for September 20 and then subtract the cumulative probability up to September 19.

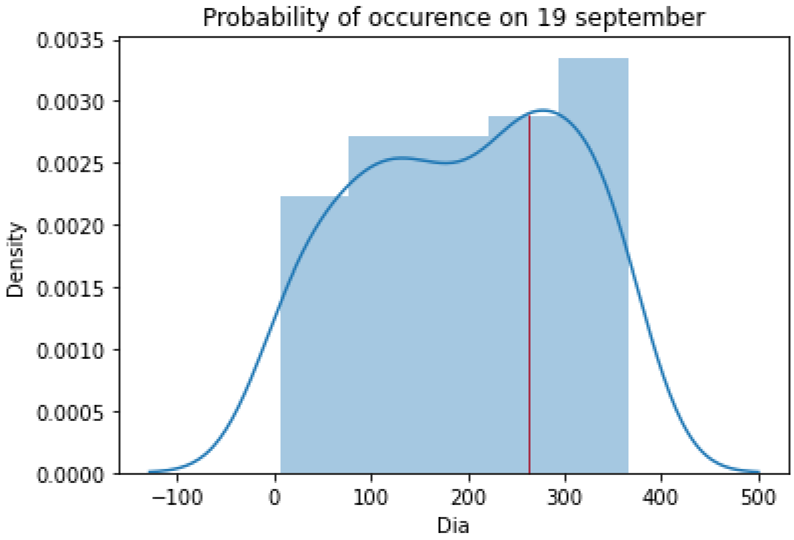

Figure 3. The red line represents the probability of a large earthquake occurring on September 19. Note that the associated probability for September 19 appears to be approx. 0.0030 (lower than its actual value*), but this is because the days on the x-axis are grouped into 5 bins in the histogram instead of the 8 bins used in figure 2. Even the PDF looks more compact.

The calculations show a probability of occurrence of a large earthquake on September 19:

$$P(X = 1) = 0.009137^{*}$$

Finally, to calculate the probability that 3 large-magnitude earthquakes have occurred in Mexico on September 19, we do:

$$P(X = 1)^3 = (0.009137)^3 = \mathbf{7.63x10^{-7}}$$

This experimental result is quite similar the theoretical probability calculated in case 1 ($\mathbf{2.1x10^{-8}}$)**,** barely 37 times greater (only 1 order of magnitude), which is logical given that we know that an increase in earthquakes occurs during the month of September.

It is important to highlight that when experimentally estimating the distribution of earthquakes during the year, we are assuming that they occurred and were distributed in some way, and from that fact, we calculate the probabilities. That is, what we have done is empirically verify case 1.

Conclusions

- The theoretical and experimental probabilities of the studied event are similar and very low.
- Although the estimated experimental probability and the theoretical one are quite similar, the first one is a little higher given that in reality, more earthquakes occur during the month of September.

- If we want to include in the calculations the probability of the occurrence of an earthquake as a random variable independent of the date of the year, the probability is still much lower than only considering the variable date of the year.
- For this project, the KDE method was used to estimate the probability density function of the real data. Using different estimation methods may lead to different results.
- Although the probability of occurrence of these 3 events is very low, they do happen.

References

- James, M.A., and Garcia-Soto (2019). *Probability of the occurrence of two significant earthquakes on the same date (of different years) striking the same site: The Mexico City case*, Seismological Research Letters, Volume 90, Number 1, 378-386.
- Kernel density estimation, https://en.wikipedia.org/wiki/Kernel_density_estimation.