# Measuring triggering in earthquake catalogs: significance vs. coincidence

## 1. Introduction

In the *Science* article, researchers found that at sites of long-term underground fluid disposal, bursts of local seismicity are sometimes set off by the passing seismic waves from distant, much larger earthquakes. (Local means relatively close to the seismic station that's doing the recording.) The stresses carried by seismic waves are relatively tiny – only a few kilopascals (kPa), compared to the thousands of kPa that are released in an earthquake – so they can only set off an earthquake if the fault is already very close to failure. This is evidence that the fluid injection is weakening surrounding faults and leaving them primed for failure, which we call "critically stressed."

But how do we know that the earthquakes were "triggered" in the first place, and didn't just occur by coincidence? In this exercise, we'll use the earthquake catalogs produced in the *Science* article and use statistics to tell us whether to believe the connection between local triggered earthquakes at injection sites and the seismic waves from large distant earthquakes occurring on the other side of the planet.

In the first step, we'll count the number of earthquakes in some time window before and after large, distant earthquakes. This will show us whether the local earthquake rates increased, decreased, or stayed the same after the distant "trigger" earthquakes. This is the data on which we will base our conclusions. In the second step, we'll use hypothesis testing to determine whether we should consider the raw data significant, or whether the observations could result from random fluctuations in the earthquake rate (coincidence).

### 2.1 The earthquake catalog

The earthquake catalog can be found in the accompanying Excel spreadsheet. Each page of the spreadsheet contains an earthquake catalog for a different seismic station (e.g. TA.130A) at a different location (e.g. Snyder). There are two catalogs each for Snyder and Prague (rhymes with "Craig") from two separate seismic stations, operating at different time intervals.

The columns of the earthquake catalog are labeled in the header row.

**A: Matlab date** – this is the date and time of the earthquake in a format recognized by the software program Matlab. It is defined (more or less arbitrarily) as the decimal number of days since January 0, 0000. This format gives us a handy way to compute differences in time between events.

**B-G: year/mo/day/hour/min/sec:** The date and time of the earthquakes in a format recognized by humans.

**H: S-P time** – the time in seconds between the arrival of the P (primary) wave, and the S (secondary) wave. This time can be used to measure the distance to the local earthquake. For these earthquakes, the distance is about 8 km per second of S-P time. For example, if the S-P time is 4.5 seconds, the distance is about 4.5 x 8 = 36 km.

**I: Magnitude** – the Richter magnitude of the local earthquake, estimated from the amplitude of the seismic waves.

The major 'trigger' earthquakes are also included in the catalogs, although they are not located nearby the seismic stations.

If you know how to use Excel, you can reproduce the figures in the paper by making a scatter plot of Magnitude vs. Matlab date, or use the DATE function in excel to use the year/mo/day columns. For this exercise, we'll just count the triggered earthquakes directly from the spreadsheet.

## 2.2 Counting triggered earthquakes

Let's compare the number of earthquakes in a small time period after the trigger to the average rate before the trigger. We want to use a relatively small window after the trigger to make sure the local earthquakes are really triggered, but a slightly longer window beforehand to average out random fluctuations in the rate.

Use the earthquake catalogs to count the number of earthquakes in the 10 days before each trigger earthquake, and the 2 days after each trigger earthquake (these quakes are in **bold text** in the catalogs). Fill in table 1 below, which has the fluid injection sites in the rows (Snyder, Prague, and Trinidad), and the three trigger earthquakes in the columns (Chile, Tohoku, Sumatra).

If you find a trigger earthquake that is repeated in two catalogs with the same injection site, use the catalog that has a larger total number of earthquakes in the catalog for that time period (the more "complete" record).

**Table 1**

| | Chile | | Tohoku | | Sumatra | |
|---|---|---|---|---|---|---|
| Number of local earthquakes: | 10 days before | 2 days after | 10 days before | 2 days after | 10 days before | 2 days after |
| **Snyder** | 3 | 0 | 1 | 4 | 0 | 0 |
| **Prague** | 0 | 33 | 0 | 0 | 0 | 2 |
| **Trinidad** | 1 | 4 | 4 | 0 | 2 | 0 |

From the numbers in Table 1, we can see a pattern emerge. The 'before' columns should on average have 5 times the number of earthquakes in the 'after' column, because the 'before' time window is 5 times as long. This doesn't seem to be the case… Which of the sites would you identify (on a preliminary basis) as likely showing triggered bursts of seismicity?

Answer: Tohoku at Snyder, Chile and Sumatra at Prague, and Chile and Sumatra at Trinidad.

## 2.3 Calculating significance of the rate changes

The next step is to show that the rate changes we observe are really significant, and unlikely to have resulted from chance coincidence. This is more commonly referred to as 'testing' our hypothesis.

### Testing a hypothesis

If you haven't encountered hypothesis testing before, don't worry – you'll be able to use the table that follows to compute statistical significance without understanding all of the theory behind it. Nevertheless, we'll go through the theory briefly, because hypothesis testing is the foundation of scientific argument. The basic strategy of hypothesis testing is to compute the probability that the pattern you observed is the result of random chance. The "random chance" model is usually called the "null hypothesis." If the probability is very low (say less than 1%), we call the observation significant.

### The null hypothesis for triggered earthquakes

As we mentioned before, if the earthquakes occurred randomly in time with respect to the trigger earthquake, they should appear 5 times more frequently in the 10 day 'before' window than in the 2 day 'after' window, because the former is 5 times longer than the latter. Our null hypothesis is therefore that there should be 5 times fewer earthquakes in the ‘after’ window than in the ‘before’ window.

This is a special case of a “binomial” process, where each observation can only have one of two outcomes – each ‘random’ earthquake lands either in the ‘before’ box or the ‘after’ box. Lucky for us, binomial processes are famously well-studied (flipping a coin is an example), and we have a simple equation for how likely it is to get exactly $k$ earthquakes in the ‘after’ box, out of $n$ total earthquakes.

## The binomial probability function

We won’t derive the binomial probability function here, but the equation looks like this:

$$\Pr(K = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \qquad \text{(Eq. 1)}$$

This is the probability of observing exactly $k$ earthquakes in the ‘after’ box, given $n$ earthquakes in total. The parameter $p$ is the probability of landing in the ‘after’ box rather than the ‘before’ box. In this case $p = 2/(10+2)$; the length of the ‘after’ time window divided by the sum of the ‘before’ and ‘after’ time windows.

Using Equation 1, we can produce a table of probabilities as a function of the number of earthquakes in the ‘after’ box and the total number of observed earthquakes (the sum of the earthquakes in the ‘before’ and ‘after’ boxes). Keep reading if you want more details of how the probabilities are actually calculated. Otherwise, skip ahead to table 2 to compute the significance of our observed rate changes.

## The cumulative binomial probability function

Equation 1 gives the probability of getting exactly $k$ earthquakes in the ‘after’ bin, but what we really want is the probability of getting $k$ or more earthquakes in the ‘after’ bin. To get this number we use a shortcut. Since the individual probabilities of all the different possible outcomes have to add up to 100%, we can take the sum of the probabilities of getting *fewer* than $k$ earthquakes (called the cumulative probability), and subtract that number from 1.

$$\Pr(K < k) = \sum_{m=0}^{k-1} \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \qquad \text{(Eq. 2)}$$

$$\Pr(K \geq k) = 1 - \Pr(K < k) \qquad \text{(Eq. 3)}$$

Equation 3 is used to produce Table 2, using a value of $p = 2/12$.

## 2.4 Identifying the significant triggers

Using Table 2, look up the probabilities of getting the earthquake counts that you entered into Table 1, given our null hypothesis that the earthquake occurrence is totally random. Enter these probabilities into Table 3. Notice that according to the model you are 100% likely to observe at least zero earthquakes in the 'after' box

**Table 2: Probability chart for the null hypothesis using windows of 10 days before and 2 days after**

| Number 'after' | number 'before' + number 'after' | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** |
| **0** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **1** | 0.17 | 0.31 | 0.42 | 0.52 | 0.60 |
| **2** | - | 0.027 | 0.074 | 0.13 | 0.20 |
| **3** | - | - | 0.0046 | 0.016 | 0.035 |
| **4** | - | - | - | 0.00077 | 0.0033 |
| **5+** | - | - | - | - | ≤0.00013 |

.

**Table 3: Probabilities for earthquakes from Table 1**

| | **Chile** | **Tohoku** | **Sumatra** |
|---|---|---|---|
| **Snyder** | 1.00 | 0.00077 | 1.00 |
| **Prague** | <0.00013 | 1.00 | 0.027 |
| **Trinidad** | 0.00077 | 1.00 | 1.00 |

We are now ready to identify the instances where the triggered rate changes are truly significant. Commonly, scientists set a probability threshold for significance that is around 0.01 or maybe 0.05. These are referred to as 99% confidence and 95% confidence, respectively (can you see why?).

Using the 99% confidence threshold (0.01), which earthquakes in Table 3 had significant triggering at each of the three fluid injection sites?

Are your findings consistent with the findings in the *Science* article?