Tres sismos
@Stacy_Nguyen_, septiembre 2022

Como ya todo mundo sabe, volvió a haber un terremoto de una cierta magnitud (7.7) en 19 de septiembre, el tercero en años recientes de magnitud 7+ (mayor o igual a 7). Muchos se han preguntado la probabilidad de que esto ocurriera y algunos han ofrecido respuestas numéricas. Presento la mía con algunas variaciones que pueden ser útiles para alguien enseñando cursos introductorios a la probabilidad y la programación.

Para los datos, voy a usar la base de datos del Sismológico Nacional. Se remonta a 1900 pero es evidente que el nivel de detalle ha ido cambiando conforme avanzó el siglo XX, así que solamente empleo el subconjunto de eventos a partir de fechas más recientes.

Respuesta analítica

La probabilidad p de que haya un sismo 7+ en un día dado es (asumiendo un sismo al día máximo):

p = no. días con sismos 7+ / número días = 45/26562 = 0.0016941

(contando de 1950) de acuerdo al Sismológico. Esta p es la misma para cada fecha, incluso el 19 de septiembre (usaré “fecha” para referirme a la combinación día-mes).

La probabilidad de que a ese sismo le sigan otros dos de magnitudes similares en la misma fecha de los siguientes 37 años después del primer sismo, (para 38 en total, como 1985-2022) están dados por una función binomial:

Bin(k=2; n=37, p=0.0016941) = 0.001801

si exigimos exactamente dos sismos más para juntar 3, o:

Bin(k>1; n=37, p=0.0016941) = 0.001837

si aceptamos más de tres sismos como parte de la secuencia de tres sismos. La probabilidad conjunta de que acontezcan 3 sismos en la misma fecha en un lapso de 38 años es por tanto el producto de las probabilidades de que ocurra un sismo en una fecha arbitraria por la probabilidad de 2 sismos más en la misma fecha en los subsiguientes 37 años:

p(3 sismos de magnitud 7+ en 38 años) = p*Bin(k=2; n=37, p=0.0016941) = **3.05 millonésimos**

o:

p(3 sismos de magnitud 7+ en 38 años) = p*Bin(k>1; n=37, p=0.0016941) = 3.11 millonésimos

si aceptamos una secuencia de 3 o más sismos en 38 años.

El mismo cálculo puede extenderse a otras combinaciones de parámetros.

Respuesta con código

El primer método es el que aplicaría cualquier persona que hubiera llevado un curso introductorio de probabilidad. Otra manera de atacar el problema es posible con elementos básicos de programación.

La idea es la siguiente: cargar la BD completa en una matriz y contar las apariciones de secuencias requeridas. La frecuencia con la que aparezcan las secuencias debería aproximar las probabilidades reales. Como dije antes, la BD comienza en 1900 y contiene unos 260 mil elementos. El problema es que el registro de sismos es muy desigual, incrementándose con el paso del tiempo. Por consiguiente, posiblemente la precisión no sea uniforme, sobre todo de los sismos de menor magnitud.

Corrí el código comenzando a partir de 3 fechas para ver el efecto (redondeado a enteros) del número de secuencias de exactamente 3 sismos de al menos la magnitud indicada en 38 años.

Año	1900			1950			1975		
m	#	p	r	#	p	r	#	p	r
4+	11638	10	8	11545	0	0	11408	0	0
5+	1507	239	75	1414	193	74	1348	72	62
6+	306	49	6	213	29	6	170	9	3
7+	86	0	1	45	0	1	31	0	1

m = magnitud
# = no sismos registrados de esa magnitud entre el año dado y 1985-09-22
p = secuencias predichas por el cálculo analítico
r = secuencias reales en el registro del sismológico

Es claro cómo, conforme avanzamos en el tiempo, la previsión analítica mejora notablemente. Una arruga: el tamaño de la “ventana”, i.e. el número de años que nos ocupa (1985-2022), es de 38. Los registros del sismológico llegan hasta 2022-09-22, lo que quiere decir que la última fecha que podemos contar en el cálculo es 1985-09-22: no han transcurrido suficientes años para hacer un juicio sobre las fechas posteriores porque aún no hay registros 38 años después. El cálculo toma eso en cuenta.

Llegado este punto, podríamos haber dado el cálculo por terminado pero me di cuenta que en realidad hay una manera aún más sencilla de verificar el cálculo (si se tiene la solución analítica) o directamente de obtenerlo con código.

El problema hasta ahora es la falta y variedad cambiante de los datos en el registro. ¿Por qué no tomar los datos que consideremos más fiables para calcular las probabilidades de que ocurra un sismo a una magnitud arbitraria y simular directamente la ocurrencia de sismos, y de ahí contar las secuencias con las características que se deseen? Eso fue lo que hice y obtuve una mejor aproximación, con unas pocas líneas de código. Esa es la mejor respuesta al problema, que no se me ocurrió hasta el tercer intento.

(Hint. Ayuda pensar en un calendario con un formato alternativo: en vez de agrupar por año, hacerlo por día, i.e. en un calendario de mil años, los primeros mil registros son del 1 de enero de los años 0..999, seguidos por 1000 registros del 2 de enero, etc.)

De ahí obtenemos, una vez más para secuencias de exactamente 3 sismos de magnitud 7+ en 38 años (corriendo 10 mil experimentos de cien mil días cada uno), tomando 1975 como fecha de referencia para el cálculo de las probabilidades de ocurrencia de un sismo en un día.

Listo varias posibilidades de acuerdo a cómo se interprete la pregunta original, dado que hay varias maneras de hacerlo.

(1) Secuencias tales por cada 100 mil días (273.8 años): 0.3346
(2) Secuencias predichas entre 1975-01-01 y 1985-09-22: 0.0131 (hubo cero)
(3) Probabilidad de que un día sea el comienzo de una secuencia tal: 3.35 millonésimos
(4) Probabilidad de que un día con sismo sea parte de una secuencia tal: 1 cienmilésimo
(5) Probabilidad de que un día, con o sin sismo, sea parte de una secuencia tal: 1.27 diezmilésimos

Creo que después del último sismo del 19 de septiembre con dos antes atrás, cuando la gente pregunta “¿qué posibilidad había de eso?”, lo que intuitivamente están preguntando es: dado que hubo dos sismos de magnitud 7+ en esa fecha, llevando la cuenta hasta 37 años atrás, ¿cuál es la posibilidad de que un tercero ocurriera en esa fecha este año? De acuerdo a (4), la probabilidad de que un día con sismo sea parte de una secuencia tal es de un cienmilésimo, pero eso incluye al primer, segundo y tercer sismos en la secuencia. Si nos concentramos en la probabilidad de que un día tuviese *el tercer* sismo en la cadena, **la probabilidad es ⅓ de eso, o 3 millonésimos** (esa sería mi respuesta con código al ejercicio, la misma que la respuesta analítica).

Nota 1. Si exigimos que el tercer sismo ocupara el lugar 38 en dicha cadena, la probabilidad es aún menor, aunque no sé si eso tenga la gente en mente cuando pregunta. Analíticamente, el resultado se obtiene tomando *p*Bin(k=2; n=36...) *p* en vez de *p*Bin(k=2; n=37...)*. Usando código, la adaptación es directa.

Nota 2. Obviamente, las probabilidades varían dependiendo de su año de referencia.

Ejercicio 1: calcular la probabilidad cuando en la cadena tres sismos en 38 años, el último de la secuencia tiene un sismo de la magnitud requerida (i.e. dos sismos repartidos en los años 1-37 y uno obligado en el año 38).

Ejercicio 2: Es posible que un día ocupe parte de una secuencia (digamos la segunda) y otra (digamos la primera) en una secuencia. Como hice el cálculo, cada ocurrencia cuenta por separado. Calcular la probabilidad contando como 1 ocurrencia independientemente de cuántas secuencias forme parte.

Ejercicio 3. Calcular las probabilidades cuando la secuencia tiene *n>=3* sismos y no exactamente 3.

Paradojas

Comienzo con un ejemplo sencillo. Imaginemos un tablero cuadrado de 100 filas y 100 columnas (i.e. 10 mil celdas) sobre el que vamos a tirar un darlo. Imaginemos también que el juego está construido de modo que no hay modo de que al tirar el dardo, caiga fuera del tablero.

Una pregunta elemental sería: ¿cuál es la probabilidad de que un dardo le atine a una casilla cualquiera, si tiramos a ciegas de modo que aproximadamente todas las casillas tengan la misma probabilidad de que el dardo les caiga? La respuesta es 1/10,000. Tanto pregunta como respuesta tienen sentido y ésta puede producirse inmediatamente. Si a continuación tiramos el dado, no nos parece para nada sorprendente que caiga en algún lado, aún con dicha pequeña probabilidad.

Consideremos otra situación aparentemente similar pero que termina en algo muy diferente. Supongamos que sin hacer preguntas ni cálculos, tiramos un dardo al azar y cae en alguna casilla. Después, nos preguntamos por la probabilidad de que hubiera caído ahí. La respuesta es la misma: una en 10 mil. Si no lo habíamos meditado antes, nos parecerá sorprendente. Alguien incluso dirá: es rarísimo que haya pasado eso, ¡debe de haber alguna explicación!

Cambió muy poco pero nuestra reacción fue muy diferente. Creo que la diferencia central entre ambas reacciones es que en la primera pasamos de la información sobre la situación a la realidad. La mente, por así decirlo, construye primero un escenario mental (“puede caer en cualquier lado por igual”) y luego, efectivamente, el dardo cae en cualquier lado. En el segundo caso, en cambio, primero nos enfrentamos al hecho y de ahí andamos hacia atrás para crear el marco en el que situar lo sucedido. La respuesta postula un hecho improbable, lo que ofende nuestra intuición, y la mente demanda una explicación.

Hay que distinguir entre hechos improbables (en el sentido probabilístico) e improbables en el sentido de extraños o sobrenaturales. Un sismo con muy baja probabilidad (digamos, uno en diez mil días) es bastante improbable en un día arbitrario, pero algo probable al cabo de mil días, medianamente probable al cabo de 5 mil, muy probable entre 8 mil, etc. La probabilidad de un fenómenos, después de todo, es una medida de su frecuencia y nada más (al menos si son “frecuentistas”).

Improbable, por tanto, no es sinónimo de “raro”. Dependiendo de cómo se construya la improbabilidad, tampoco quiere decir “no va a suceder”. La realidad es que suceden hechos improbables (en el sentido probabilístico) todo el tiempo. La direccionalidad de los sucesos vuelve a ser relevante. Si se tiene la liberta de seleccionar *hechos que ya sucedieron*, es muy sencillo identificar sucesos improbables. Por ej., puedo asomarme por la venta y ver un pájaro en el árbol. Puedo comenzar: ¿qué probabilidad hay de que me asome por la ventana y vea un pájaro? Digamos, 1/25. Me miro los calcetines. Continúo: ¿qué probabilidad hay de que me asome por la ventana, vea un pájaro y traiga calcetines azules? Digamos, (1/25)*(1/20). Miro el calendario. ¿Y en día jueves? (1/25)*(1/20)*(1/7). Etcétera.

(La práctica llamada *data dredging* se desaconseja por esta razón.)