Ley de los grandes números

Alberto Benavides
1 de diciembre de 2020

1. Introducción

La ley de los grandes números afirma que mientras un experimento aleatorio se repita n veces, el promedio de los resultados $\frac{X_1,X_2,...,X_n}{n}$ de ese experimento se aproximará a su valor esperado $E(X)$ conforme n se aproxime a $+\infty$.

Los ejemplos más usados para demostrar la ley de los grandes números consisten en lanzar una moneda al aire o la tirada de un dado de seis caras.

Para el lanzamiento de una moneda, el $E(X) = 0.5$. Los lenguajes computacionales permiten experimentar computacionalmente mediante la generación de valores pseudoaleatorios[1]. En R[2] se puede realizar un experimento que simule mil lanzamientos de monedas para, de ellos, calcular la media, con el procedimiento `mean(sample(0:1, 1000, replace=TRUE))`. Con una semilla fija `set.seed(33)`, se obtiene por resultado $0.507 \approx E(X)$. En este caso, el `0:1` simula los valores que puede tomar la variable aleatoria X: `0` para cara y `1` para cruz.

En el caso de mil tiradas de un dado de seis caras, se puede simular con `mean(sample(1:6, 1000, replace=TRUE))`, donde el valor obtenido con la misma semilla es $3.554 \approx E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$.

Una representación gráfica de la parte en la que n se aproxima a $+\infty$ consiste en graficar las medias conforme n crece. Para el ejemplo de la tirada del dado de seis cara en la que se registran las tiradas en un rango $i \in [1, 10\ 000]$, se logra mediante el código 1.

Código 1: Medias experimentales de las tiradas de un dado de seis caras

```
1  medias <- c()
```

[1] https://es.wikipedia.org/wiki/Generador_de_n%C3%BAmeros_pseudoaleatorios
[2] https://www.r-project.org/

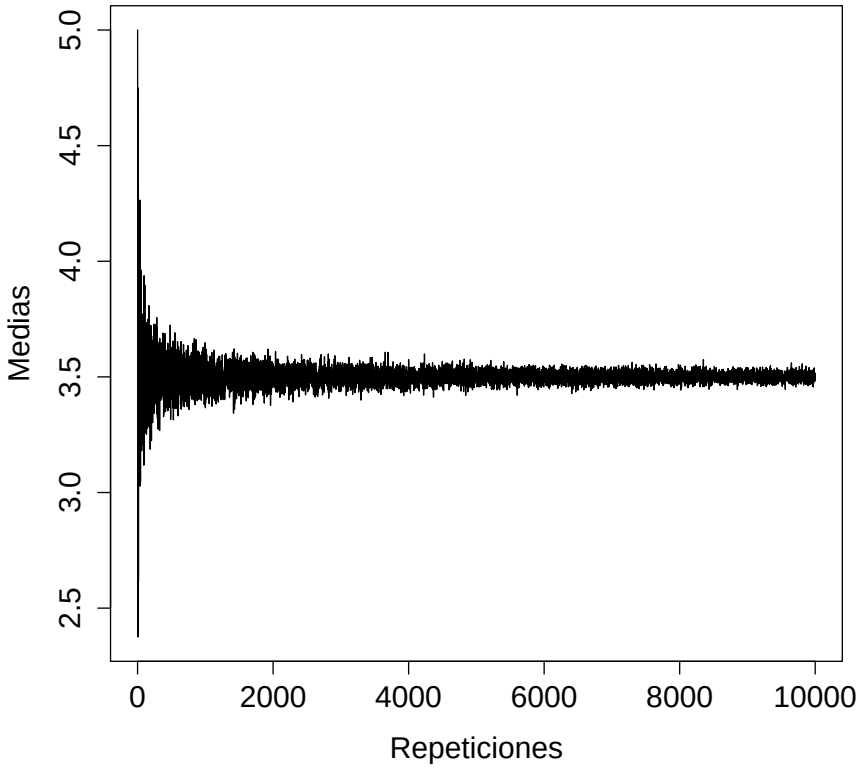

Figura 1: Series de tiempo diferenciadas de casos registrados de enfermedades y concentraciones de partículas de 10 PM en el AMM durante 2017.

```
2   for (i in 1:10000){
3       medias <- c(medias, mean(sample(1:6, i, replace=TRUE)))
4   }
```

Las medias de cada i número de experimentos se almacenan en una variable cuyos valores, mostrados en la figura 1 (p. 2), permite ver cómo converge al $E(X)$ el resultado de las medias conforme el número de repeticiones del experimento se hace más grande.

2. Problema

Actualmente, investigo la relación que hay entre los contaminantes del aire y los reportes de enfermedades por parte de los centros de salud, estudio delimitado al Área Metropolitana de Monterrey durante el año 2017. Los datos de contaminantes para ese año se obtuvieron del Sistema Integral de Monitoreo Ambiental (SIMA) [1]. Un ejemplo de los datos obtenidos puede verse en la tabla 1 (p. 3), donde se constata que existen entre los datos la fecha en que fueron registrados, una de las trece estaciones que hizo el

Tabla 1: Muestra de mediciones capturadas por los sensores del área metropolitana de Monterrey.

Fecha	Estación	CO	NO	...	Válida
19-Aug-16 16	Centro	0.43	1.90	...	1
23-Mar-97 0	Sureste	1.23	1.25	...	1
21-Oct-11 0	Norte	2.16	40.00	...	1

registro, los valores de ciertos contaminantes. Además, hay una columna llamada válida, Elisa Schaeffer[3] agregó para diferenciar las mediciones inválidas conforme a lineamientos que el SIMA también proporcionó. Si la medición es válida, se asigna un valor de 1, mientras que si es inválida en alguno de los valores reportados, se le asigna un valor de 0.

Uno de los problemas que se ha tenido con estos datos es la cantidad de errores en medición que presentan y una de las preguntas que se plantean por esta situación es qué estrategias se pueden implementar para mejorar la precisión de los sensores. En esta tarea se presenta una respuesta que utiliza la ley de los grandes números para encontrar un número de mediciones que se deberían hacer a partir de la media de valores inválidos registrados para saber si un cambio o reparación en los sensores mejora su certeza.

3. Solución

Los errores de medición detectados por Elisa Schaeffer se agruparon por año y estación de monitoreo, obteniendo por resultado registros como los que se muestran en la tabla 2 (p. 3).

Tabla 2: Cantidad de errores agrupados

Año	Estación	Errores	Datos por grupo
1999	Noroeste	40	8760
2003	Suroeste	0	8760
2000	Centro	35	8784

La estación con más errores de medición, con un total de 41.47 %, es la Suroeste, cuya ubicación geográfica se muestra en la figura 2 (p. 4).

Por medio de la ley de los grandes números se puede calcular la cantidad de mediciones que se deben hacer para saber, con un 95 % de confianza, que el porcentaje de errores de medición se mantiene. En el caso de la estación Suroeste, se sabe que la probabilidad de obtener una medición errónea es $\mu = 0.4147 = p$, de donde $\sigma^2 = p(1-p) = 0.2427$. Ahora, se puede definir un error de medición $\epsilon = 0.02$, lo que quiere decir que se probará si una medida es errónea con probabilidad entre $[0.4147 - 0.02, 0.4147 + 0.02]$. Por la ley

[3] https://elisa.dyndns-web.com/

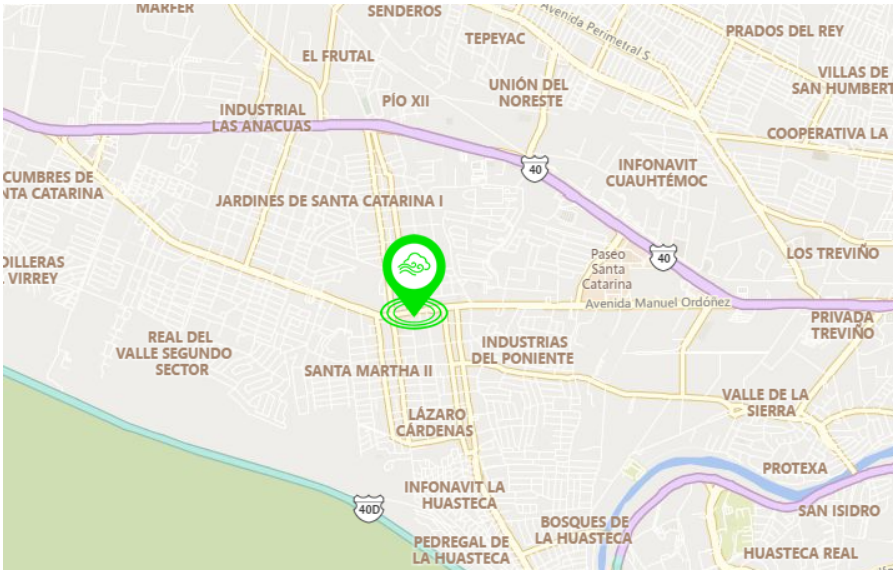

Figura 2: Mapa que muestra la ubicación de la estación de monitoreo Suroeste con un marcador de posición verde. Imagen obtenida de la página del SIMA[1].

de los grandes números, se tiene que

$$P[|\bar{X} - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2},$$

al sustituir

$$P[|\bar{X} - 0.4147| > 0.02] \leq \frac{0.2427^2}{n(0.02)^2}.$$

Como se desea obtener un intervalo de confianza de 95 %, entonces se debe cumplir

$$\frac{0.2427^2}{n(0.02)^2} = 0.05,$$

por lo que

$$n = \frac{0.2427^2}{(0.05)(0.02)^2} = 2945.1645 \approx 2945.$$

Así, se podría recomendar realizar $n = 2945$ mediciones y comparar el porcentaje de errores encontrados, para saber si alguna modificación en los sensores ha modificado la certeza de sus mediciones.

Referencias

[1] Sistema Integral de Monitoreo Ambiental [SIMA] Nuevo León (2020), «aire.nl.gob.mx | Home», `http://aire.nl.gob.mx/`, [Accedido 14/may/2020].