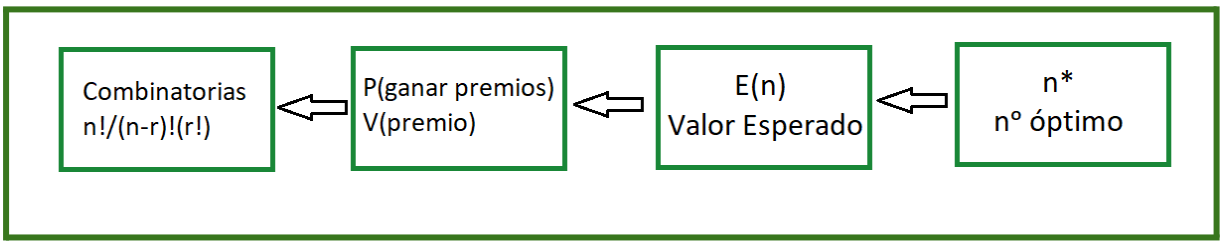
Ignacio Pulido Ruiz
Diego Horcajo Abal

El juego de la lotería: Contexto del reto

El sorteo de la Lotería de Navidad es un clásico de nuestra época que se ha convertido en parte del ritual navideño en muchos hogares. ¿Quién no ha comprado alguna vez un boleto, ya sea de forma individual, compartido, en familia o entre compañeros de trabajo? Esta tradición, que tiene lugar cada 22 de diciembre, viene acompañada por la expectación de esa mañana, en la que los niños del colegio de San Ildefonso cantan los números ganadores y, por tanto, es legítimo preguntarse cuál es la probabilidad real de resultar agraciado en este sorteo y si realmente merece la pena participar. Surge la siguiente pregunta: ¿cuántos boletos debería comprar para maximizar el retorno de mi inversión? Estudiaremos la existencia de un punto óptimo en el que uno maximiza sus ganancias **esperadas** según la posibilidad de salir premiado.

Nuestro reto consistirá en encontrar el número de papeletas a adquirir para garantizar un máximo de rentabilidad de tal forma que necesitaremos definir una función objetivo de valor esperado, lo cual requiere, en primer lugar, determinar 1) las probabilidades asociadas a cada premio y 2) la remuneración de los mismos. Este ejercicio se nos complica cuando buscamos una expresión general para un número n de boletos, ya que dichas probabilidades y escenarios cambian conforme varía nuestra variable independiente (número de papeletas adquiridas):



En el flujo anterior se observa el procedimiento necesario para llevar a cabo el estudio del número óptimo de boletos comenzando por delimitar las combinaciones ganadoras posibles para un número determinado de papeletas, como explicaremos en próximas secciones.

Presentamos una visión más específica de los principales premios en la tabla siguiente, centrándonos en los siete premios más destacados, ya sea en términos de ganancias o de probabilidad asociada.

Premio	Cantidad de Números Premiados	Probabilidad Asociada (%)
El Gordo (1er Premio)	1	0.001
2º Premio	1	0.001
3º Premio	1	0.001
4º Premio	2	0.002
5º Premio	8	0.008
Reintegro	9,999	10
La Pedrea	1,794	1.794

En el caso del reintegro, nos tropezamos con aquellos boletos premiados cuya última cifra coincide con la del número ganador. Dado que se sortean 100,000 series en total, formados por cinco cifras, contamos con 10 posibles opciones para cada uno de los primeros cuatro dígitos (desde la primera hasta la cuarta cifra). No obstante, para la quinta cifra, la restricción es que esta debe coincidir con la del número ganador, lo que limita nuestro número de combinaciones ganadoras para el reintegro. Así, se generan 10 * 10 * 10 * 10 (10,000) series posibles. Sin embargo, es necesario restar una combinación: la del propio número ganador de El Gordo, ya que si coincide, no se estaría ganando el reintegro, sino el primer premio. Por lo tanto, el número de boletos premiados con el reintegro es de 10^4 - 1 = 9,999 papeletas premiadas con el reintegro.

Mi primer boleto

Realmente, considerando que aproximadamente un 12% de los boletos serán premiados de una forma u otra (ya sea con un premio mayor o menor), no parece tan improbable que nuestro número resulte premiado en algún año. Además, si compramos más de un boleto con combinaciones distintas, aumentaremos nuestras opciones de ganar.

Asumiendo esta probabilidad fija del 12% (frente a un 88% de probabilidades de no ganar nada) y definiendo la posibilidad de ganar algo como $1 - p(perderlo\ todo)$, se puede calcular que sería necesario jugar un mínimo de 5 números distintos para asegurarse de ganar algo en alguno de los boletos con una probabilidad mayor al 50%. ¡No está nada mal para empezar!

Por otro lado, si restringimos este cálculo al número de boletos necesarios para asegurarnos, con una probabilidad mayor al 50%, de obtener alguno de los grandes premios, el resultado asciende a aproximadamente 5300 números. Eso sí, este cálculo no garantiza una ganancia neta positiva, ya que habría que descontar el coste de la inversión inicial (20 euros por número). Lo único que asegura es ganar algún premio con probabilidad superior al 50%.

Por este motivo, y para hacer el análisis más interesante, no solo nos basaremos en probabilidades, sino también en el retorno esperado, considerando el coste inicial en los siguientes apartados.

Valor Esperado de nuestro juego

Para analizar los posibles resultados de nuestra lotería, habiendo definido previamente los distintos premios con sus series y sus probabilidades asociadas, procederemos a estudiar la ganancia esperada por cada juego. Supongamos que adquirimos un único boleto de lotería. El espacio muestral de este experimento abarca diversas posibilidades: ganar el primer premio, el segundo, recibir una pedrea, o incluso no obtener ningún premio en absoluto. En este contexto, surge una pregunta crucial: ¿cuánto podemos esperar ganar al comprar un boleto?

Calcular la probabilidad de ganar uno de los premios es bastante sencillo, ya que está estrechamente vinculada al número de series asignadas a cada premio. Por ejemplo, solo 1 de cada 100,000 series tiene el primer premio, mientras que hay 8 series que contienen el quinto premio, lo que se traduce en una probabilidad de 8/100,000 para este último. Además, es fundamental considerar el costo inicial de comprar un boleto, que en España es de 20 euros.

De acuerdo con la teoría del valor esperado, debemos considerar la suma ponderada de todos los casos posibles por sus probabilidades y la ganancia neta correspondiente a cada uno. Formalizamos mediante la siguiente ecuación para el cálculo del valor esperado:

$$\mathbb{E}(X) = \sum_{i=1}^{n} P(X_i) \times (G_i - C)$$

Donde $P(X_i)$ es la probabilidad de obtener el premio i-ésimo, G_i es la ganancia bruta asociada a dicho premio, y C es el coste del boleto.

El problema se vuelve más complejo cuando deseamos calcular el valor esperado de nuestra ganancia neta para un número general de boletos n. En otras palabras, queremos asociar un valor esperado al juego de la lotería para cualquier cantidad de boletos adquiridos. Para ilustrar la complejidad del problema, comencemos particularizando para $n = 2$. Con dos boletos, los posibles escenarios incluyen ganar solo el primer premio, ganar solo el segundo, o ganar el tercero, entre otros. Sin embargo, también aparecen alternativas intermedias, como ganar el primer y el segundo premio, o ganar el segundo y el tercero, así como combinaciones más avanzadas como ganar el primero y el quinto premio simultáneamente.

Para n=3, la cantidad de escenarios posibles se incrementa considerablemente, lo que nos lleva a una mayor generalización y a la consideración de alternativas adicionales. Surge entonces una pregunta: ¿Como de significativos son estos contextos intermedios? Por ejemplo, ¿es razonable pensar que una persona podría ganar simultáneamente el segundo y el tercer premio? Si eso ocurriera, ese individuo sería extraordinariamente afortunado… ¡más que jugar a la lotería, parecería estar dictando su propio destino! Sin embargo, como veremos a continuación, desarrollaremos desde cero la fórmula para calcular el valor esperado para cualquier cantidad de boletos n. Esta fórmula será capaz de abarcar todos los escenarios posibles, desde no ganar nada hasta la improbable, pero posible, eventualidad de ganar el primer, segundo y tercer premio al mismo tiempo.

Para entender mejor la ecuación que determina el valor esperado en nuestra lotería, es esencial definir las variables clave que utilizaremos:

- S: Número total de series emitidas en la lotería.
- n: Número de boletos comprados.
- i: Índice que representa el premio específico (donde i varía de 1 a 7, ya que hay 7 premios).
- V_i: Valor monetario asociado al premio i.

La probabilidad asociada a cada escenario en la lotería la determinaremos utilizando el operador de combinatoria, que permite calcular el número de formas en que se pueden seleccionar ciertos boletos entre el total disponible. Para un número n de boletos comprados, la probabilidad de que el premio esté entre tus boletos se puede expresar de la siguiente manera:

$$\left(\frac{100000 - \sum_{i=1}^{7} S_i}{n-1}\right) \cdot S_i = \frac{(100000 - \sum_{i=1}^{7} S_i)!}{(n-1)! \cdot \left((100000 - \sum_{i=1}^{7} S_i - (n-1))!\right)} \cdot S_i$$

El numerador representa las combinaciones de $n - 1$ boletos no ganadores, dado que ya hemos seleccionado un boleto ganador (asumiendo que un individuo compra distintos números sin repetir). Este término cuenta cuántas formas hay de elegir $n - 1$ boletos de los $100000 - \sum S_i$ boletos restantes que no tienen el premio i ni ningún otro premio. En el caso del primer premio, i sería igual a 1 y S_i tendría el valor de 1, puesto que solo hay una serie (número) correspondiente al primer premio. Multiplicamos por un factor correctivo de S_i para tener en cuenta aquellos premios con más de una serie premiada (e.g. 8 quintos premios). Lo que realmente estamos haciendo es aislar el escenario en el que ganamos un único premio, excluyendo la posibilidad de ganar múltiples premios. Esta simplificación nos permite analizar de manera más clara el número de combinaciones posibles y, como hemos visto previamente, la probabilidad de ganar más de un premio es extremadamente pequeña o despreciable, lo que justifica que podamos ignorarla para realizar el cálculo. Si, por el contrario, incluyéramos este escenario en la ecuación, no tendríamos una forma precisa de calcular la ganancia, ya que implicaría una combinación de múltiples premios. Esta situación sería demasiado variable y, por tanto, imprecisa.

$$\binom{100000}{n} = \frac{100000!}{n!(100000-n)!}$$

Este término cuenta todas las formas posibles de seleccionar n boletos de un total de 100000 y nos permitirá delimitar el denominador de nuestra expresión de probabilidad.

$$V_i - C = V_i - 20n$$

El cálculo anterior nos da el valor real que se obtiene después de descontar el costo del boleto del valor del premio ganado (multiplicado por el número de boletos comprados). Esto es esencial para evaluar la ganancia efectiva de participar en el juego. Al juntar todas estas partes, obtenemos la siguiente expresión para el valor esperado:

$$E(n) = \left(\sum_{i=1}^{7} \frac{\binom{100000 - \sum_{j=1}^{7} S_j}{n-1}}{\binom{100000}{n}} \cdot S_i \cdot (V_i - 20n) \right)$$

Esta expresión final nos permite calcular el valor esperado de la ganancia neta al comprar n boletos de lotería, considerando todos los premios posibles y sus respectivas probabilidades. Al iterar sobre cada uno de los 7 premios, podemos evaluar de manera integral las expectativas de ganancia al participar en el juego. Ahora analizaremos el último escenario y el más probable, en el que no ganamos nada. En este caso, la pérdida será linealmente proporcional al precio del boleto, dado que por cada boleto comprado se perderá esa cantidad. La probabilidad de este escenario se calcula como el número de combinaciones posibles de seleccionar n boletos de entre las series no premiadas, dividido por el número total de combinaciones posibles de todos ellos. Esta probabilidad refleja la alta posibilidad de no obtener un premio, mientras que la pérdida se escala en función del número de boletos comprados.

$$E(n) = \frac{\binom{100000 - \sum_{j=1}^{7} S_j}{n}}{\binom{100000}{n}} \cdot (-20n)$$

Poniendo todos nuestros resultados juntos:

$$E(n) = \left(\sum_{i=1}^{7} \frac{\binom{100000-\sum_{j=1}^{7} S_j}{n-1}}{\binom{100000}{n}} \cdot S_i \cdot (V_i - 20n) \right) + \left(\frac{\binom{100000-\sum_{i=1}^{7} S_i}{n}}{\binom{100000}{n}} \cdot (-20n) \right)$$

La limitación principal de la ecuación anterior, como hemos mencionado, es que no contempla escenarios intermedios en los que se pueda obtener más de un premio simultáneamente con "n" boletos. Esto se debe a que en nuestra ecuación de combinatorias estamos considerando directamente la posibilidad de ganar un único premio y, a partir de ahí, calculamos el número de combinaciones con los boletos restantes (n-1) entre las papeletas no premiadas. Si incluyéramos en el cálculo las papeletas premiadas, obtendremos una probabilidad más precisa. Sin embargo, esto haría imposible asignar de forma crítica una ganancia específica a cada escenario, ya que no podríamos determinar si estamos ganando un premio, dos, tres o más de manera simultánea.

Además, para valores grandes de n, nuestra ecuación no aplica del todo ya que el denominador crece de manera factorial más rápido que el numerador, por lo que a partir de cierto punto esta expresión tenderá a 0 y no nos aporta ningún valor visual al graficarlo. Para valores pequeños de n, dicha expresión parece aproximarse a un comportamiento lineal en función de n que deducimos de las partidas de "$-\ 20n$". Para dar un paso más en nuestra gráfica de retorno esperado en función del número de boletos adquiridos, realizaremos una simulación numérica en el último apartado.

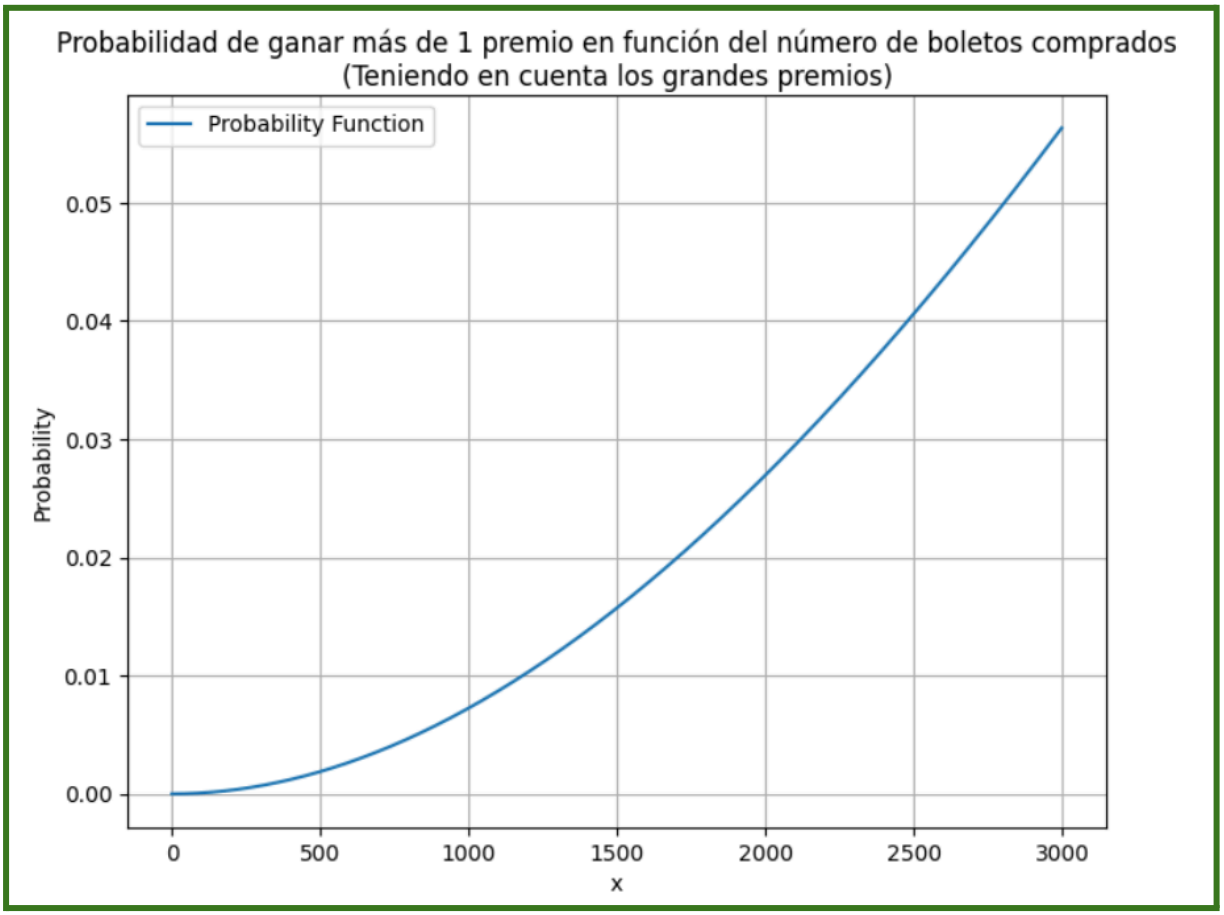
¿Es relevante el escenario de ganar más de un premio?

Como mencionamos en los apartados anteriores, hemos excluido escenarios intermedios para simplificar el análisis probabilístico. Sin embargo, cabe preguntarse si el hecho de ganar más de un premio fuese realmente un escenario relevante en términos de probabilidad. Aunque no conocemos casos recientes en los que alguien haya ganado, por ejemplo, el primer y el tercer premio, sería fascinante que esto ocurriera (y afortunados nosotros de conocer a esa persona).

Para estudiar estos casos intermedios, aplicaremos la misma lógica inicial con la que calculamos la probabilidad de ganar alguno de los premios (ya sea mayor o menor). En este caso, explicaremos la probabilidad de ganar más de un premio entre los grandes premios. La clave es partir del espacio muestral total (que tiene una probabilidad de 1) y restarle los escenarios en los que ganamos **uno o ninguno de los grandes premios**. Así, nos quedamos con la probabilidad de ganar más de un premio:

$$P(\text{ganar más de un premio}) = 1 - \left(\frac{\binom{100000 - \sum_{i=1}^{7} S_i}{n}}{\binom{100000}{n}} + \sum_{i=1}^{7} \left(\frac{\binom{100000 - \sum_{j=1}^{7} S_j}{n-1}}{\binom{100000}{n}} \cdot S_i \right) \right)$$

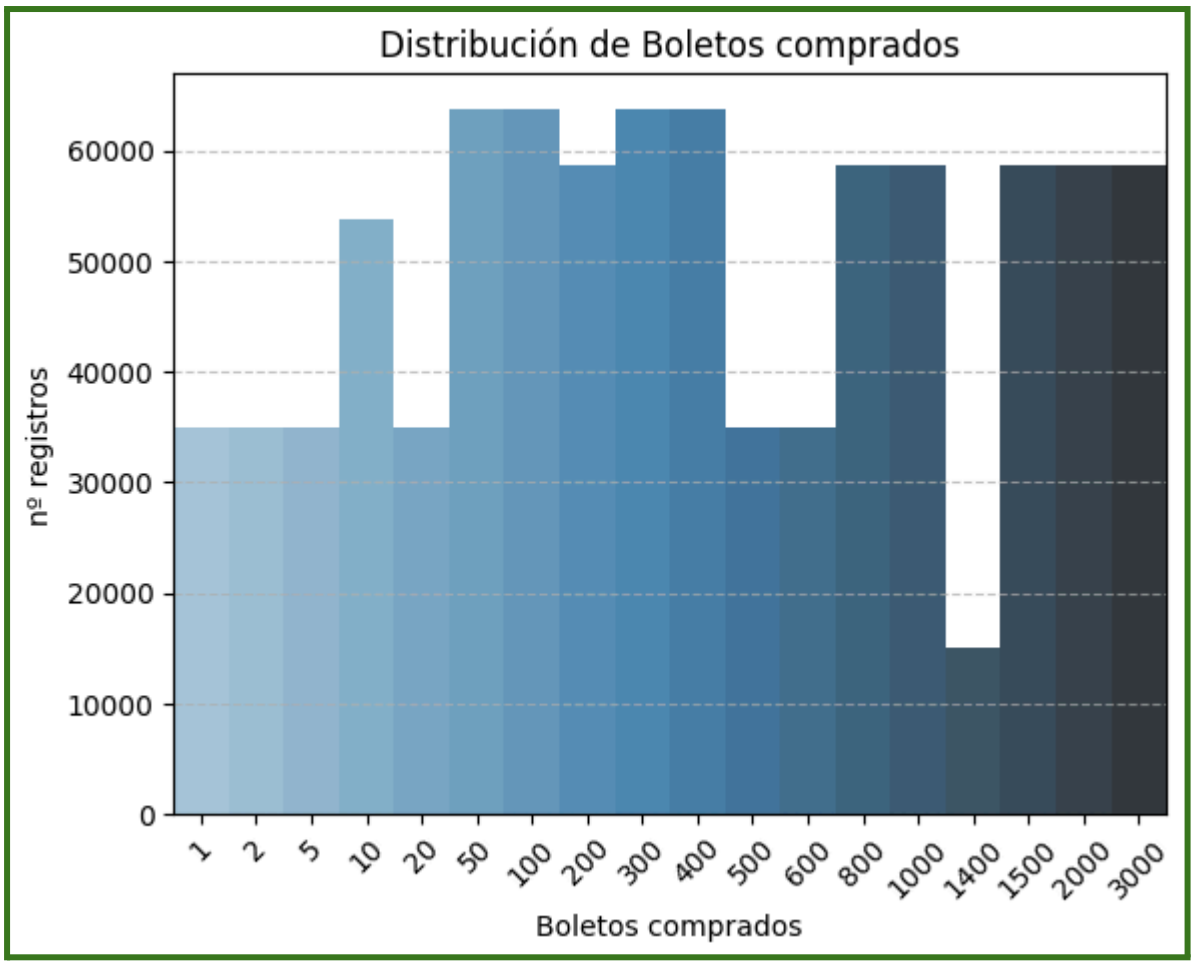
Si graficamos nuestros resultados para distintos valores de n con ayuda de nuestro simulador…



Observamos que la probabilidad de obtener dobles premios o más, incluso comprando miles de boletos, es ínfima.

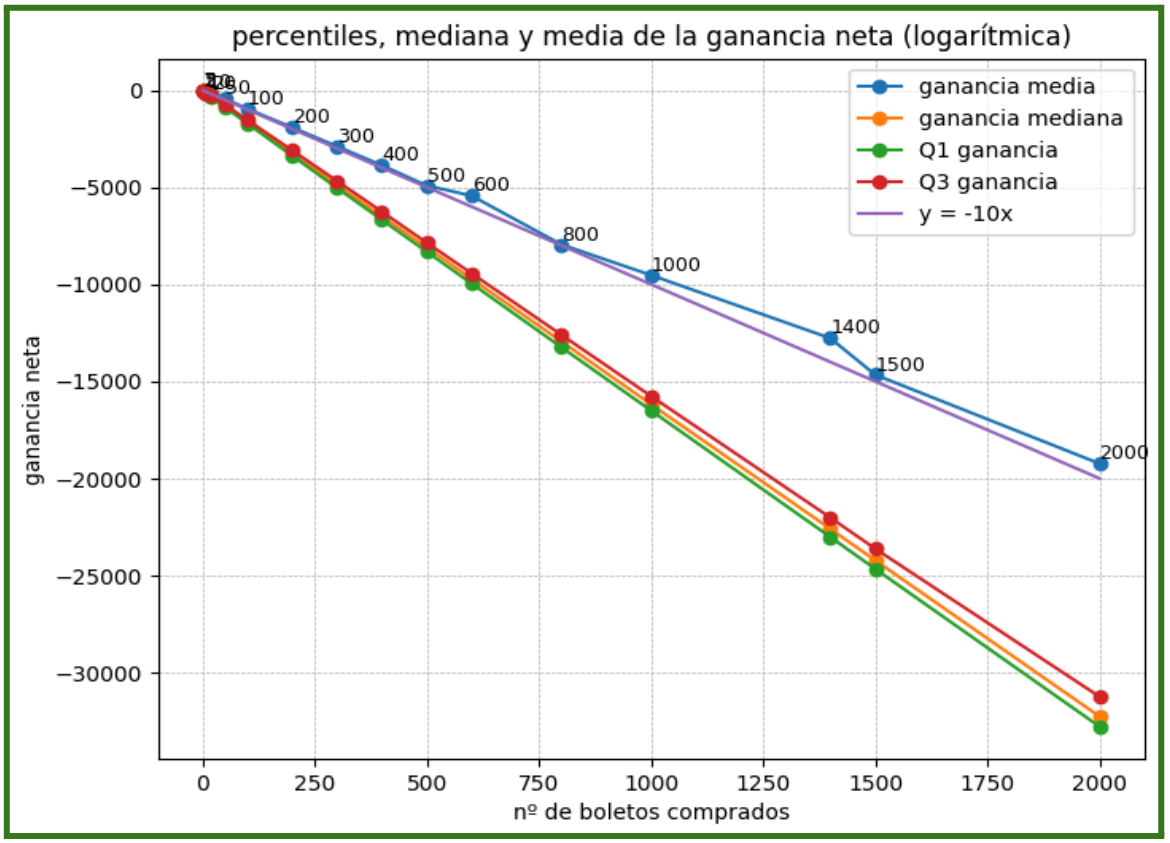
Simulador en python con medio millón de intentos

Generamos un simulador automático en el que le cargaremos nuestro escenario de Lotería de Navidad con datos sobre ganancias y probabilidades asociadas. Por cada iteración, un sujeto "i" adquirirá un número determinado de boletos y se le asignará una ganancia neta teniendo en cuenta los premios obtenidos y el coste asociado de compra de papeletas. La simulación con Python consistirá en un conjunto de muestras diferenciadas por el nº de boletos comprados por cada individuo. La distribución de muestras aparece en el siguiente gráfico:

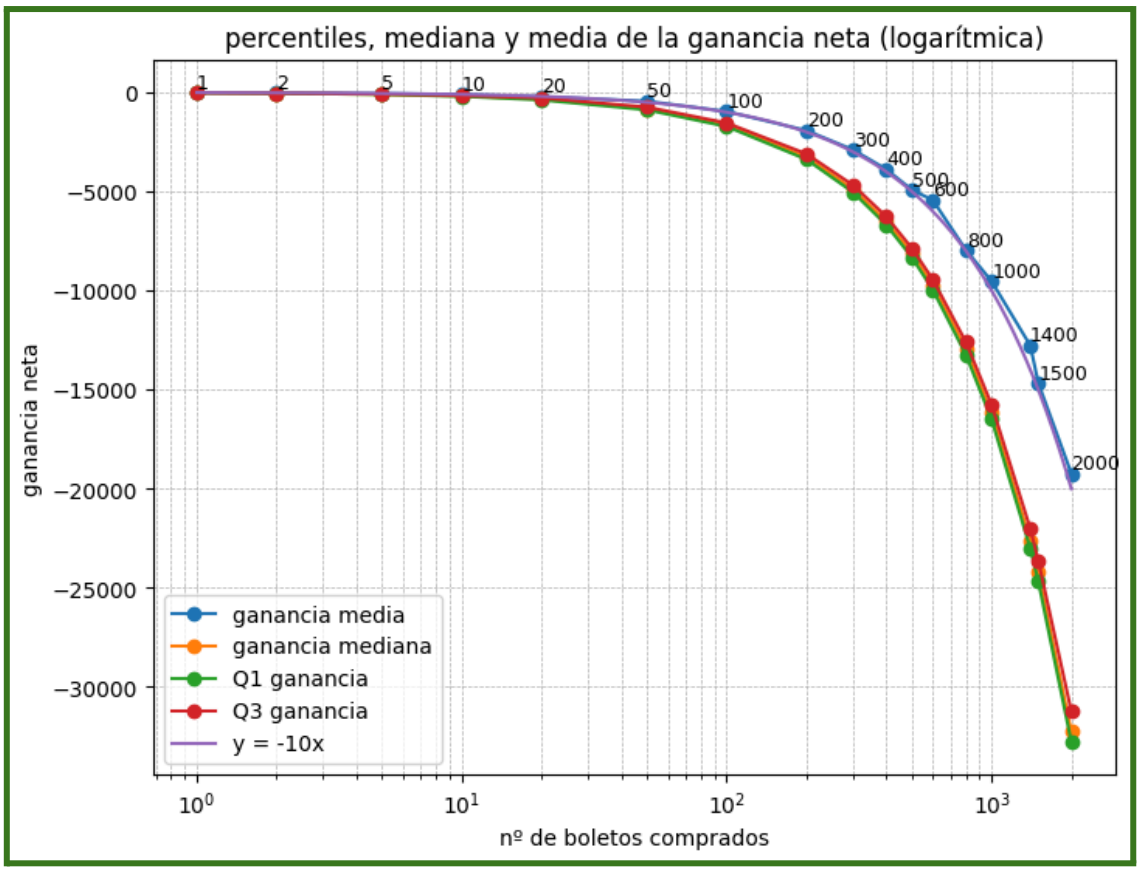


En concreto, se han realizado simulaciones comprando desde 1 boleto hasta 3000, con valores discretos intermedios. El objetivo de este experimento es doble. Por un lado se pretende extraer la evolución de la ganancia neta (es decir, la diferencia entre el beneficio obtenido por los premios y el coste intrínseco de los boletos) conforme el nº de boletos comprado aumenta, y por otro lado, ligado a esta primera visualización, la evolución de la probabilidad de obtener un retorno positivo en la inversión.

Para el primer experimento, se obtienen los siguientes resultados:



Como puede apreciarse, llama la atención la evidente linealidad entre el nº de boletos comprado y la ganancia neta esperada. En concreto, la relación es de aproximadamente -10n, donde n denota el nº de boletos comprados. Veamos esta gráfica en escala logarítmica para poder apreciar mejor los casos iniciales, no apreciables a escalas más grandes:



Deducimos que, por cada boleto adquirido, se espera perder 10 euros (de media) o, lo que es lo mismo, y teniendo en cuenta el coste inicial de 20 euros, de media se espera recuperar el 50% del coste inicial. Este resultado, hablando en medias, siempre se aleja de lo terrenal, puesto que al final existirá quien obtenga ganancias positivas, un grueso importante de participantes que lo pierdan todo y, en pocos casos, personas que sean premiadas en grandes cantidades. Otro punto que nos llama la atención es el comportamiento lineal de nuestra curva de retorno esperado. El resultado se alinea con nuestra hipótesis inicial en la que construimos la ecuación de permutaciones en la que, hasta cierto nivel "n", el comportamiento era bastante homogéneo. Recordemos que, para valores pequeños de n, dicha expresión parecía aproximarse a un comportamiento lineal en función de n que deducimos de las partidas de "$-20n$". Después de todo, observamos que, de media, cada boleto adquirido es una variable aleatoria con las mismas probabilidades de ganar un premio o perder, por lo que, al comprar "n" boletos, el valor esperado de retorno será la suma de los valores esperados de todos los boletos (y más si tenemos en cuenta que varias personas podrían compartir un mismo número).

Escenario extra:

Nos interesa determinar la probabilidad de obtener un rendimiento positivo adquiriendo "n" boletos. Por ejemplo, si compramos un único número, para poder obtener un beneficio (ingreso - coste > 0) debemos considerar la probabilidad de obtener cualquier premio cuya dotación supere los 20 euros. Si obtenemos dos boletos, entonces ese espacio muestral incluye la posibilidad de tener ambos boletos premiados con dotación mayor que 20 euros, o también la posibilidad de tener un beneficio en uno de ellos lo suficientemente grande como para cubrir el coste inicial de 40 euros. A medida que aumentamos nuestra variable de número de boletos, nuestro espacio muestral se multiplica y se vuelve muy tedioso el cálculo. Un **approach** inicial sería el siguiente:

$$P = P(ganancia > 0\ con\ 1\ boleto) \approx 1.8\% \rightarrow p(ninguna\ ganancia) = 1 - P = 98.2\%$$

Asumiendo que nuestra variable es tal que cada boleto es independiente entre sí, podemos delimitar la distribución binomial tal que:

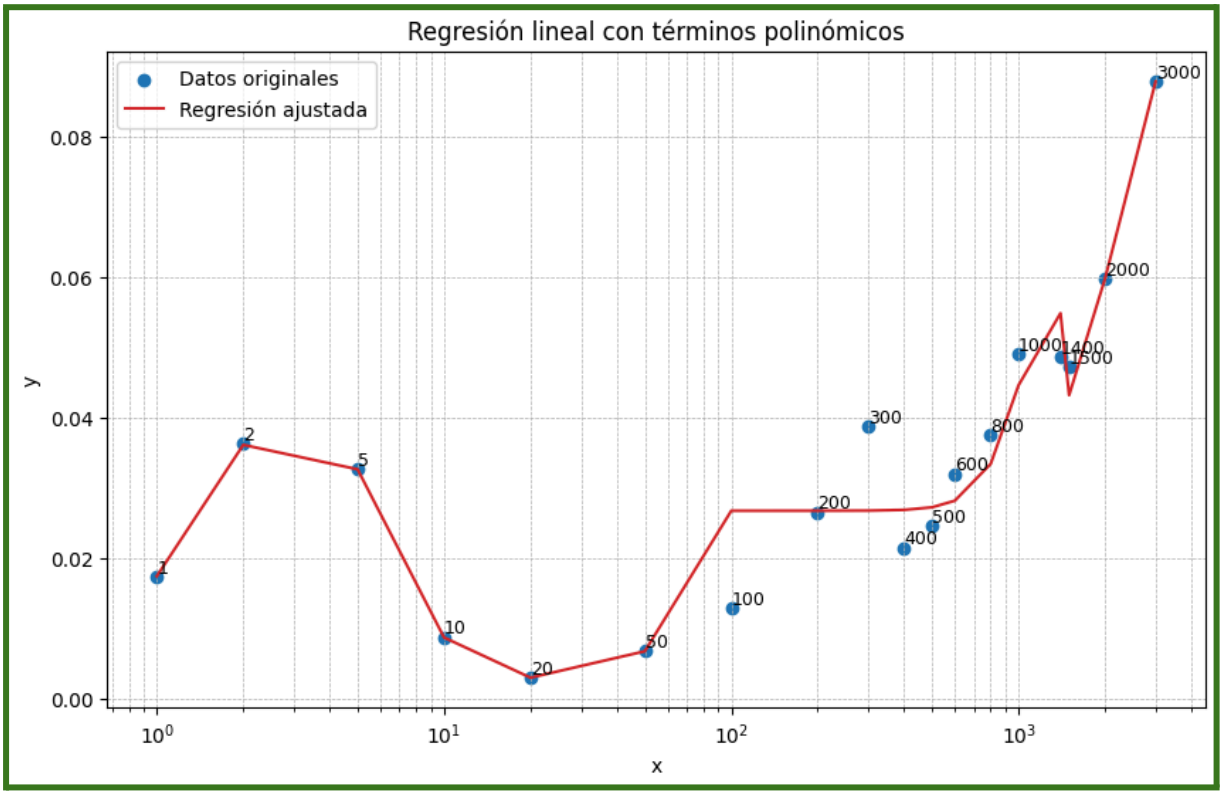
$$X \sim B(n, 0.982) \rightarrow Distribución\ de\ pérdidas$$
$$\rightarrow p(pérdidas\ con\ n\ boletos) = (0.018)^0(0.982)^n$$
$$\rightarrow p(no\ pérdidas\ con\ n\ boletos) = 1 - (0.018)^0(0.982)^n$$

De tal forma que, por ejemplo, con 10 boletos, la probabilidad de tener ganancias sería de aproximadamente del 16.6%. **Sin embargo,** uno se da cuenta de que lo que realmente estamos calculando aquí es la probabilidad de **no** tener pérdidas en todos nuestros boletos, cuya probabilidad no es la misma de tener ganancias en agregado. Es decir, estamos incluyendo en esta probabilidad todos los escenarios en los que **almenos uno** de nuestros números tiene ganancia, pero no necesariamente cubre el coste inicial. Por tanto nuestra probabilidad real será menor. Por otro lado, podemos explorar la probabilidad de tener ganancia positiva en todos y cada uno de los 10 boletos:

$$p(ganancias\ con\ n\ boletos) = (0.018)^{10}(0.982)^0 < 1\%$$

Pero, al igual que en el caso anterior, esto está asumiendo una ganancia positiva en todos y cada uno de nuestros números. Por tanto, la probabilidad real de ganancia positiva con "n" boletos sabemos que existe entre los dos escenarios que acabamos de proponer.

Se ha evaluado la probabilidad de obtener una ganancia neta positiva en el conjunto de simulaciones. Le ajustamos un modelo de regresión polinómica y observamos el comportamiento de la curva:



Los resultados arrojados muestran cómo entre 1 y 300 boletos, la probabilidad de obtener un retorno positivo de la inversión fluctúa en torno el 2-4% de probabilidad, con un valle pronunciado entre los 5 y 100 boletos comprados. La principal hipótesis ante este fenómeno apunta a una falta de simulaciones en los primeros casos: 1, 2 y 5 boletos, pues como se ha visto en la gráfica del comienzo de esta apartado, la cantidad de simulaciones por cada boleto es dispar, teniendo menos en los primeros casos. Seguramente, con un conjunto más amplio o equilibrado de simulaciones, la función de probabilidad de retorno positivo tienda a suavizarse y reducir el número de picos.