Análisis de distintas estrategias en el problema Multi-armed Bandit con estadística bayesiana

Trabajo Práctico 1

Tomas Anderson Alejo Vaschetti Manuel Moresi

Tabla de contenidos

Introduccion

El humano tiene que enfrentarse a una gran cantidad de decisiones en su vida diaria. Algunas pueden ser irrelevantes, como la elección de para qué lado dormirse o cuál es el primer pie al que se le pone la media al vestirse. En cambio, otros dilemas pueden ser de una importancia superior en la que se espera un mayor uso de racionalidad. Por ejemplo, al terminar la escuela, ¿Empezar a trabajar o estudiar?, ¿Qué carrera seguir?, ¿Salir de fiesta con tanta lluvia el fin de semana?. Estos problemas de decisión que conllevan un poco más de esfuerzo pueden ser analizados de forma dicotómica, basar la elección en base a la seguridad de lo familiar y el conocimiento previo o decantarse por lo desconocido y arriesgado pero con posibles recompensas mayores. Usando los ejemplos anteriores, ¿Seguir una carrera que se adecúe a las técnicas propias o elegir una carrera de conocimientos inexplorados?,¿Quedarse en casa viendo una serie o salir de fiesta pudiendo ser una noche inolvidable o una desastrosa?. Este dilema se lo conoce como "explorar vs. explotar" en el cual el primero hace referencia a la promesa de lo desconocido y lo segundo a la seguridad de lo familiar.

En este informe se analizará una situación hipotética que plantea la disyuntiva de "explorar vs. explotar". El problema del *Multi-armed Bandit* se trata de imaginarse una situación en donde se tienen varias máquinas tragamonedas, es decir, máquinas en las que se hace girar unos rodillos con figuras en cada cara y en el caso de coincidir la figura se gana dinero, en caso contrario, no se gana nada. Para simplificar, se analizará el problema con el caso de 3 máquinas con distintas probabilidad de ganar, 0.3, 0.45 y 0.55 respectivamente, y donde jugar en la máquina no cuesta dinero. Cada día se puede jugar en una única máquina una sola vez y el objetivo será ganar la mayor cantidad de unidades monetarias en un año bisiesto (366 días) idealmente descubriendo la máquina con mayor probabilidad de éxito. Este informe busca comparar distintas estrategias que tratarán de encontrar un equilibrio en el dilema "explorar vs. explotar".

Desarrollo

Para la resolución de las incógnitas planteadas en la introducción se usarán herramientas correspondientes al análisis bayesiano donde la creencia a priori para las probabilidades de éxito de cada máquina a principio de año toma una distribución Beta$(2, 2)$. Al considerar la variable como ganar o no al jugar una máquina, esta toma una distribución Binomial$(1, p)$ por lo que la ditribución a posteriori tiene forma de Beta$(2 + 1, 2)$ si la observación fue un éxito y Beta$(2, 2 + 1)$ en el caso de perder. Por la secuencialidad de las técnicas bayesianas, esta distribución a posteriori pasa a ser la probabilidad a priori en el siguiente día, de esta manera se puede ir actualizando día a día la creencia sobre las probabilidades de éxito de cada máquina.

Se comparan siete estrategias utilizándo simulaciones de, al principio un año para cada una y luego de 1000 años para ver las posibles variaciones que pueda llegar a tener cada método

debido a la naturaleza aleatoria de las máquinas tragamonedas. Se programa una función capaz de elegir una máquina siguiendo la estrategia seleccionada, devolver un resultado y actualizar las creencias sobre las probabilidades de éxito y es la que se usará para realizar las simulaciones.

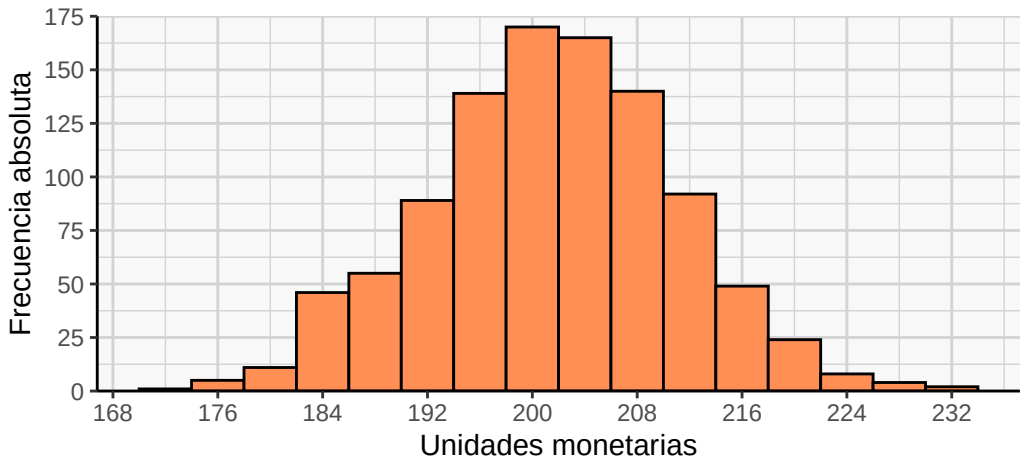
Para tener un punto de referencia con las que se pueda comparar cada método se procede a continuación el planteo de un caso irreal e idóneo en el que se consigue la mayor cantidad de éxitos en un año.

Situación ideal

La situación ideal seria conocer de antemano cual máquina es la que tiene la mayor probabilidad de ganar y jugar siempre con ella. Esta sería la mejor forma de ganar dinero, por lo que una buena estrategia tiene que asemejarse a dicha situación.

Para ver la cantidad de dinero que gana en promedio este método se simulan a 1000 personas jugando 1 año utilizándolo.

Estos fueron los resultados:



En este modus operandi una persona que juege un año se espera que va a ganar alrededor de 201.986 unidades monetarias.

En el 95% de los años simulados las ganancias totales se encontraron entre 184 y 219 unidades monetarias. Dicho intervalo, y los que seran presentados en las próximas estrategias, se

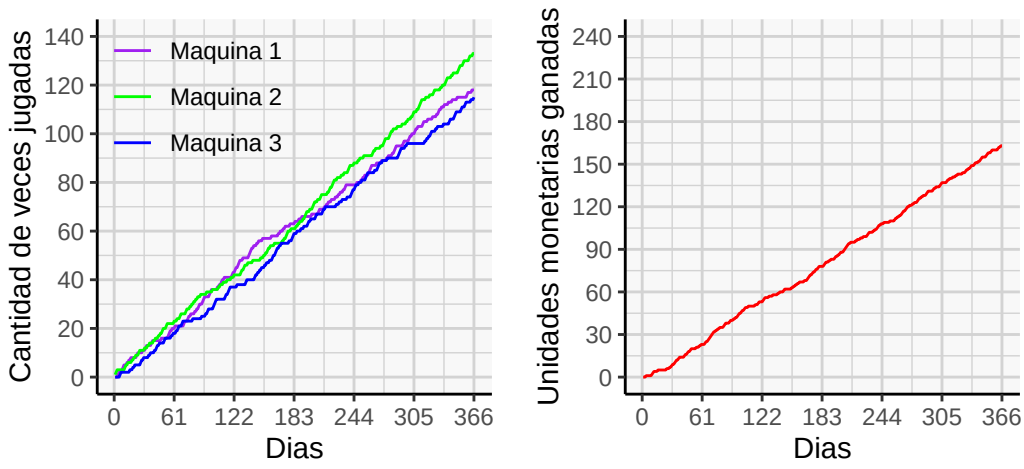
encuentran entre el 2,5% y el 97,5% de las ganancias totales del método ordenadas de menor a mayor. La amplitud del intervalo es de 35 unidades monetarias.

Para ver que tan beneficioso es cada estrategia, se comparan la cantidad de veces que se gana en promedio esperando que se acerque lo máximo posible a esta forma de jugar. Además, se contrastan las amplitudes de los intervalos propuestos.

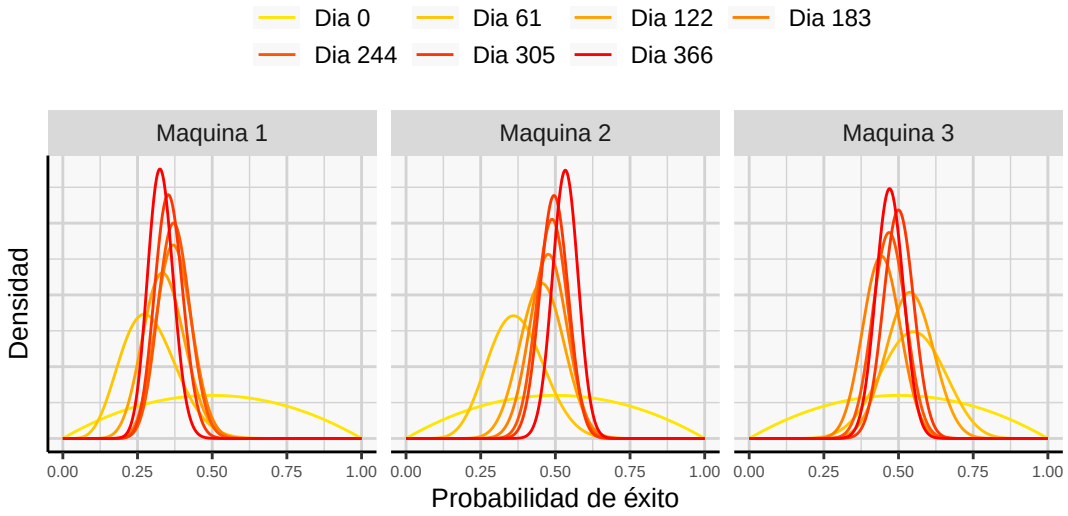
Estrategia 1: Completamente al azar

Esta estrategia consiste en elegir cada día al azar qué máquina jugar, teniendo las tres la misma probabilidad de ser selectas. Esta es la estrategia mas basica, ya que no requiere ningun tipo de informacion previa ni evoluciona en el tiempo.

Se juega un año con esta estrategia y se obtienen los siguites resultados:

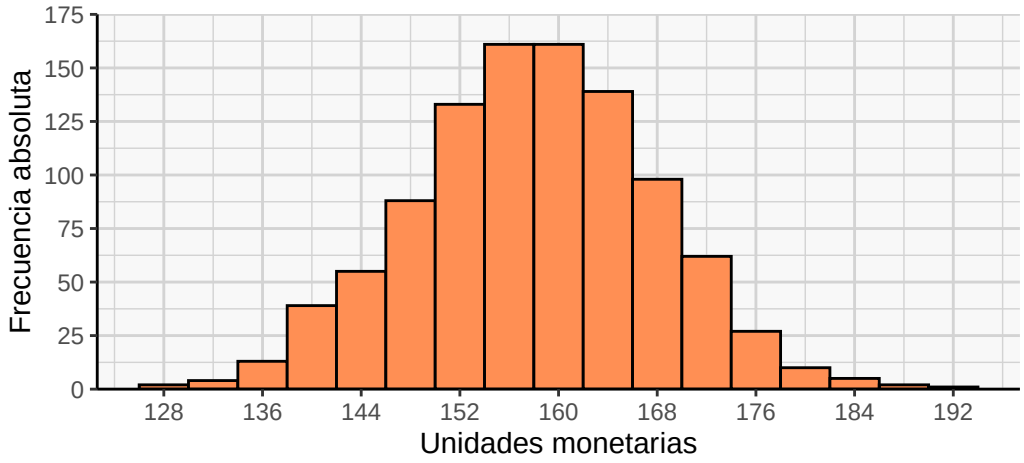


La cantidad de veces que se juegan las 3 máquinas son similares, y la diferencia entre ellas se debe puramente al azar. Este hecho impacta en la evolución de la creencia sobre la probabilidad de éxito de cada máquina, como se ve a continuación.



Al jugarse una cantidad similar de veces en cada máquina la creencia se concentra en igual medida alrededor de los valores de las probabilidades de éxito de cada una. Estos son 0.3278689, 0.5328467 y 0.4705882 para la máquina 1, 2 y 3 respectivamente.

Las simulaciones de los 1000 años arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 158.5 unidades monetarias en un año. Esto es un 21.53% menos que el caso ideal.

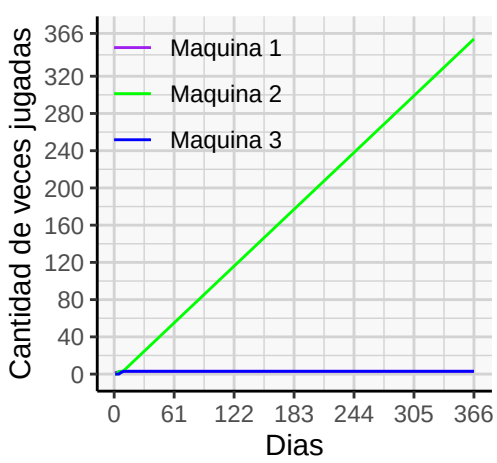
El 95% de las ganancias al final del año estuvieron entre 140 y 177. Su amplitud es de 37 unidades monetarias, siendo 0 mas grande que el del caso ideal, siendo muy parecidos entre si.

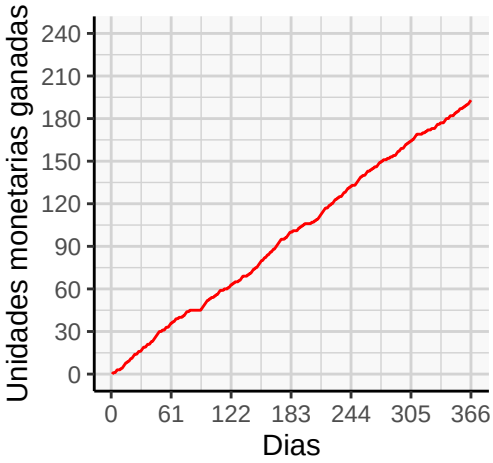
El promedio de ganancias al final del año es sustancialmente mas chico que el de la situacion ideal y los intervalos siquiera se superponen. Se puede pensar que esta estrategia no es recomendable al diferir tanto de un resultado ideal.

Estrategia 2: Greedy con tasa observada

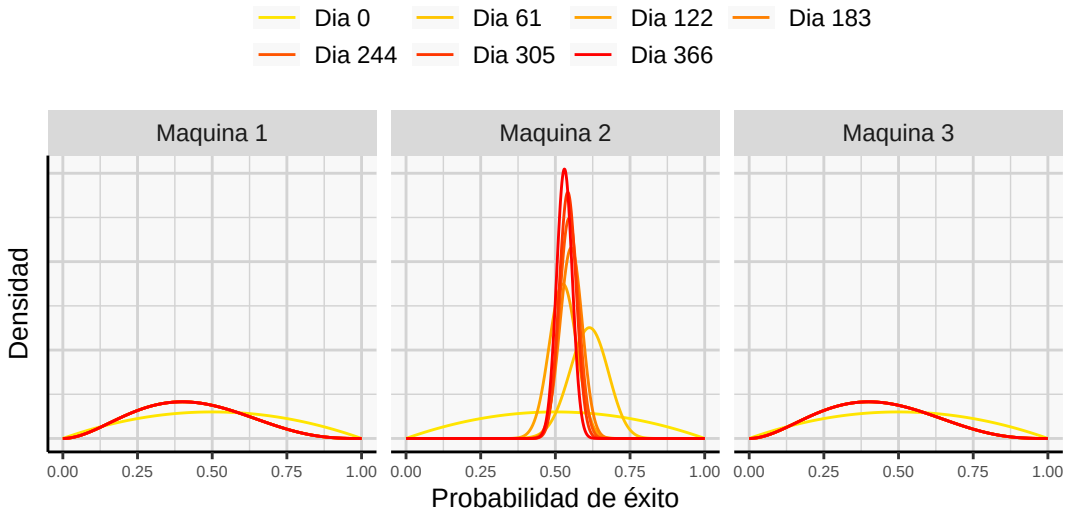
La maquina que esta estrategia selecciona en un cierto dia es aquella que tenga la mayor probabilidad de éxito observada hasta el momento. Como al empezar el año no se tiene ninguna informacion sobre la probabilidad de éxito de cada maquina, se decide jugar cada una por separado hasta que todas tengan su primer éxito. Si las probabilidades de éxito coinciden en algunas maquinas se elige una de ellas aleatoriamente.

Se juega un año con esta estrategia y se obtienen los siguites resultados:



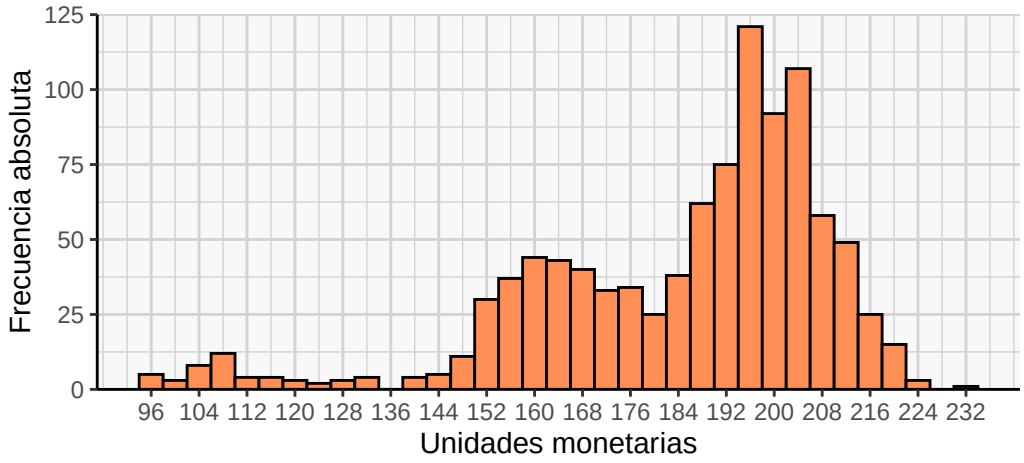


En este año se usa casi exclusivamente la maquina 2 y apenas se utilizan las otras dos. Esto pasa porque las probabilidades de exitos iniciales observadas son muy variables, por lo que la primera maquina en tener varios fracasos va a tener una probabilidad muy chica y se dejara de usar. Esto impacta a la creencia sobre la probabilidad de éxito de cada máquina de la siguiente manera.



La creencia sobre la probabilidad de exito de la maquina 2 va a ser muy precisa, al tener muchas muestras de ella, y las demas seran muy parecidas a la suposicion inical.

Las simulaciones de los 1000 años arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 185.1 unidades monetarias en un año. Esto es un 8.36% menos que el caso ideal.

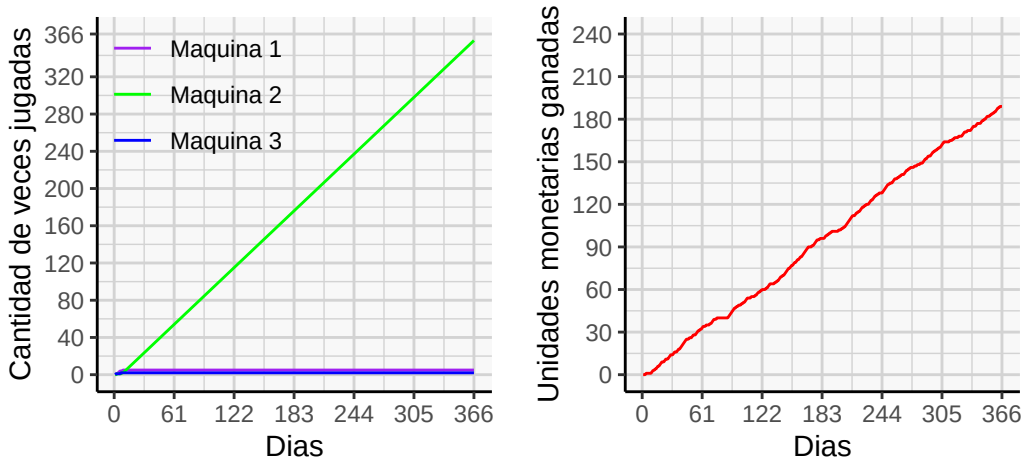
El 95% de las ganancias al final del año estuvieron entre 109 y 218. Su amplitud es de 109 unidades monetarias, siendo 3.11veces mas grande que el caso ideal.

Si bien es un metodo que en promedio gana una cantidad similar de dinero comparado a la situacion ideal, existe el riesgo de ganar mucho menos que el promedio, ya que en general la estrategia suele usar una misma maquina en casi todo el año y hay una probabilidad no despreciable de elegir la peor de ellas. Esta es la razon de porque la amplitud del intervalo es tan grande.

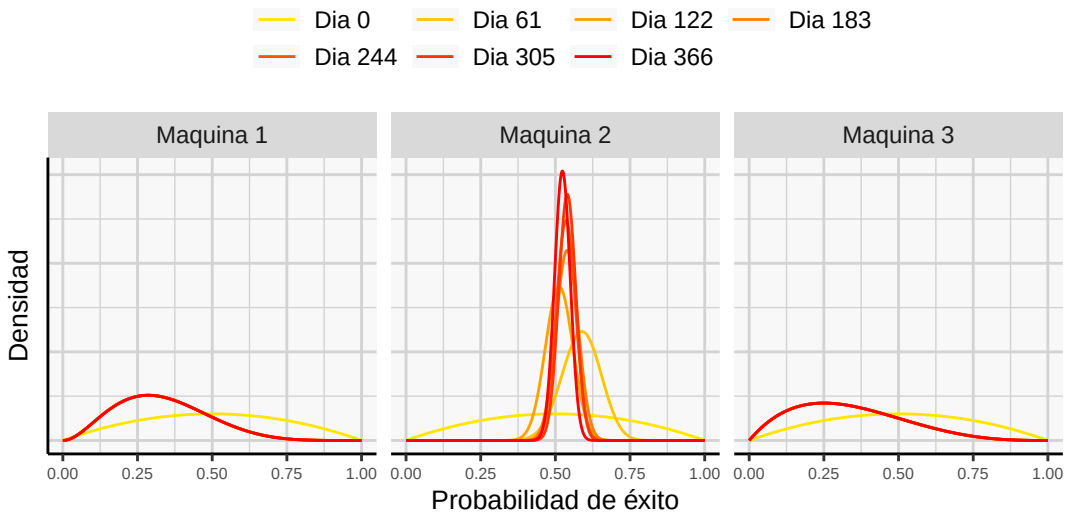
Estrategia 3: Greedy con probabilidad a posteriori

En esta estrategia, la maquina a elegir en cada dia es aquella que, hasta el momento, tenga la mayor esperanza entre las distribuciones a posteriori de las probabilidades de éxito. Si las maquinas tienen esperanzas iguales entonces se elige una de ellas aleatoriamente.

Se juega un año con esta estrategia y se obtienen los siguites resultados:

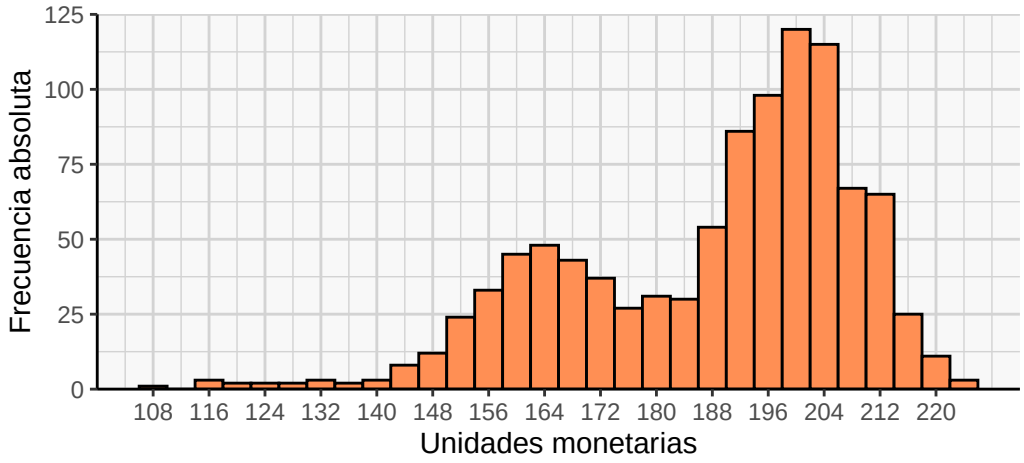


La seleccion de las maquinas en el año simulado es notablemente similar a la estrategia anterior. La unica diferencia es que las maquinas 1 y 3 se usaron un poco mas al principio del año, aunque sea dificil de ver.



Este metodo genera muchas observaciones de una maquina e ignora las demas, debido a su parecido con la estrategia anterior, por ende, la precision sobre la creencia de las probabilidades de exito sera grande para una maquina y chica en las otras.

Las simulaciones de los 1000 años arrojaron los siguientes resultados:



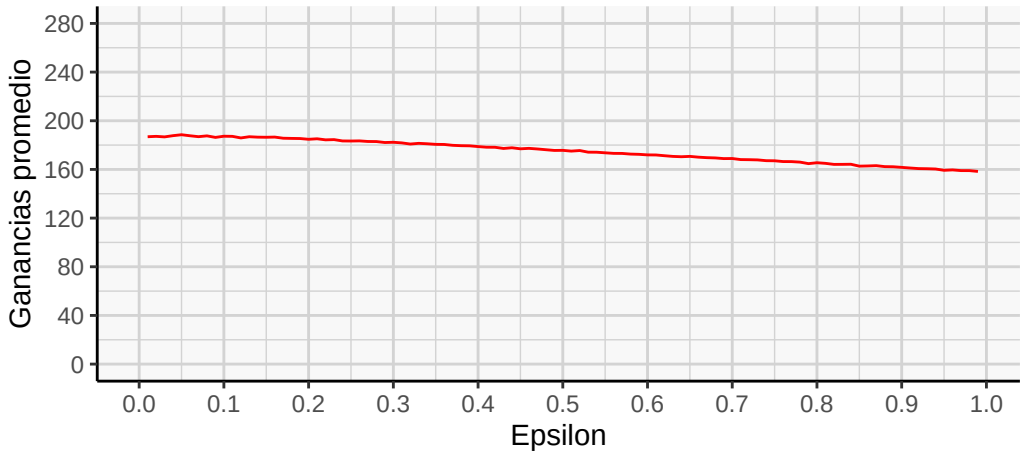
En promedio, siguiendo esta estrategia, se espera ganar 188.53 unidades monetarias en un año. Esto es un 6.66% menos que el caso ideal.

El 95% de las ganancias al final del año estuvieron entre 145 y 216.Su amplitud es de 71 unidades monetarias, siendo 2.03veces mas grande que el caso ideal.

Esta estragia sera un poco mejor que la anterior ya que la cola izquierda del histograma es menos pesada. Esto se debe a que al principio de cada año se tiene una mejor aproximacion de las probabilidades de éxito al haber utilizado la informacion de sus distribuciones a priori. Aun asi, comparada con la situacion ideal, sus ganancias son mas variables, por lo que no es muy recomendable.

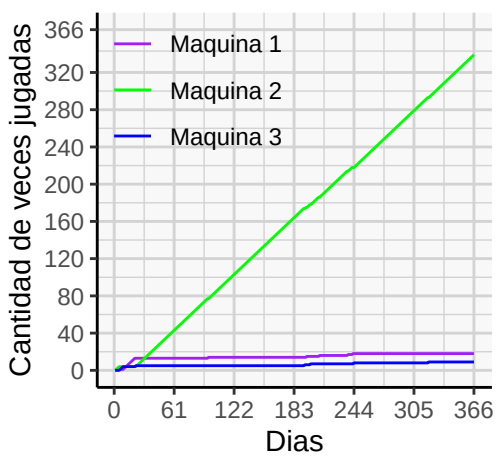
Estrategia 4: ϵ-greedy (con tasa observada)

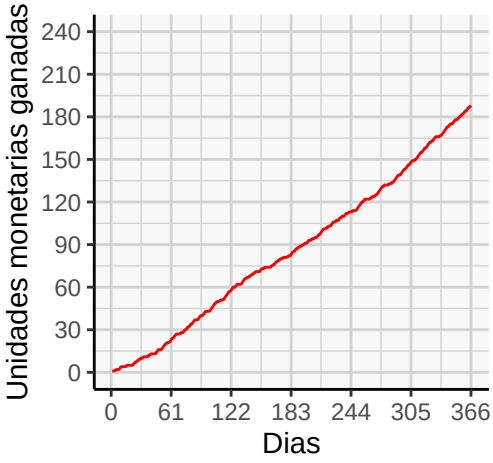
Esta estrategia es una combinacion de otras dos anteriores, la greedy con tasa observada y la completamente al azar. Se elige utlizar la primera mencionada con una probabilidad de $1-\epsilon$ y la segunda con una probabilidad de ϵ. Mientras mas cerca este ϵ de 1, se parecera mas a la completamente al azar y mas cerca este del 0 a la greedy con tasa observada. Debido a que uno decide que valor asignarle a ϵ, se realiza un analisis previo para obtener su mejor valor.



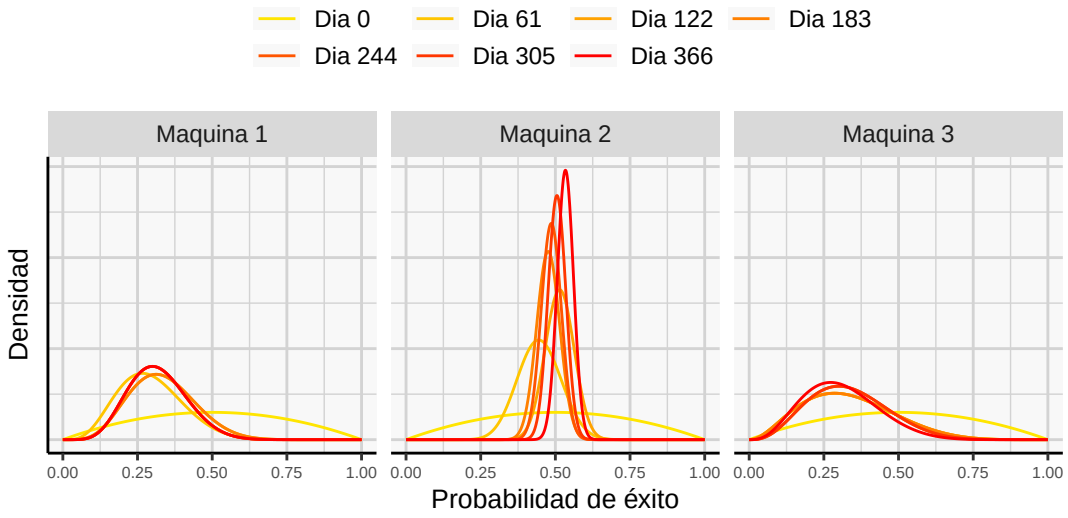
Queremos que en promedio se gane lo mayor posible. El valor que maximiza las ganancias es 0.05.

Se juega un año con esta estrategia y se obtienen los siguites resultados:



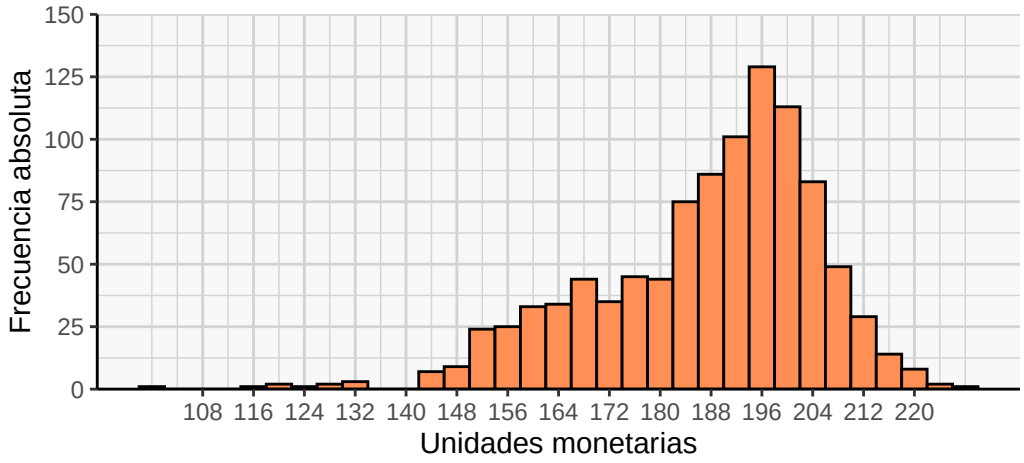


En el primer mes de juego, se decide cual es la maquina con mayor probabilidad de exito y luego se utiliza mayoritariamente una sola maquina hasta el final del año. La maquina 1 y 3 se utilizan en el resto del año solo cuando se usa la estrategia completamente al azar.



La precision de la creencia de la maquina 2 es alta, ya que es la que se utiliza con mas frecuencia. La de las demas luego de los primeros meses no cambia mucho.

Las simulaciones de los 1000 años arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 187.67 unidades monetarias en un año. Esto es un 7.09% menos que el caso ideal.

El 95% de las ganancias al final del año estuvieron entre 149 y 214. Su amplitud es de 65 unidades monetarias, siendo 1.86veces mas grande que el caso ideal.

El problema de la greedy con tasa observada es, a veces, no utilizar mayoritariamente la maquina con mayor probabilidad de exito. El usar la estrategia completamente al azar esporadicamente, permite explorar las probabilidades de exito de las maquinas para que al utilizar la greedy con tasa observada, esta seleccione la mejor de ellas con mayor regularidad. Comparado con la greedy con probabilidad a posteriori, su ganancia promedio y variabilidad son levemente menores. Aun asi, su intervalo es demasiado amplio en comparacion a la situacion ideal.

Estrategia 5: Softmax

En esta estrategia, para elegir la maquina a jugar primero se obtiene la probabilidad de exito observada π_i de cada maquina y se les aplica la siguente formula:

$$\Pr(i) = \frac{e^{\pi_i/\tau}}{\sum_{j=1}^{3} e^{\pi_i/\tau}}$$

Siendo un parametro de “temperatura” que controla el grado de exploracion. Luego se elige la maquina i con probabilidad Pr(i).

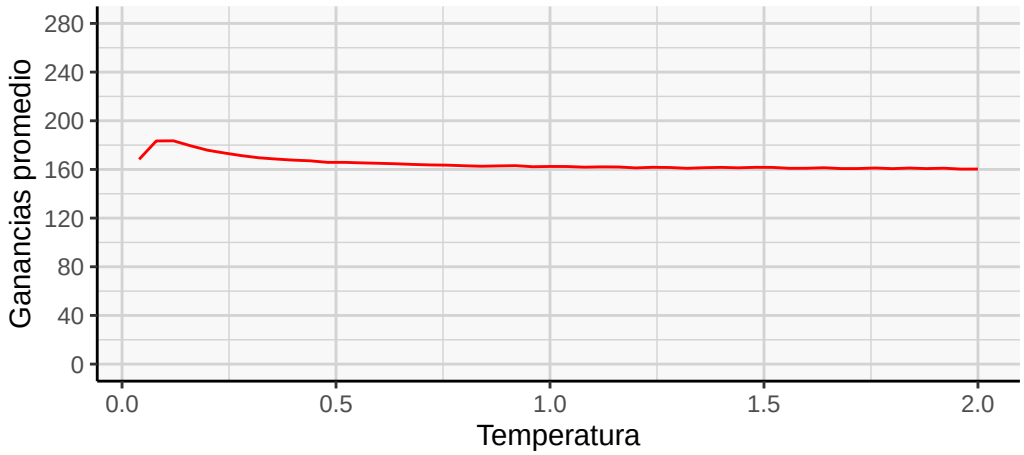
Se observa en la siguiente tabla las tasas de las probabilidades de exito reales de las maquinas para ver que les ocurre cuando varia.

	Maquina 2 / Maquina 1	Maquina 2 / Maquina 3
Originales	1.83	1.22
Temperatura = 1/25	557.77	13.12
Temperatura = 1	1.28	1.11
Temperatura = 25	1.01	1.00

Cuando la temperatura crece, la diferencias entre las probabilidades de exito de cada maquina disminuyen mientras que las diferencias aumentan cuando la temperatura disminuye. Cuando la temperatura es alta, se exploran todas las maquinas, mientras que cuando es baja, se explota la que hasta el momento sea la que tenga la mayor probabilidad de exito.

Para ver que temperatura seleccionar se grafican las ganacias promedio en 1000 años para distintos valores de la temperatura.

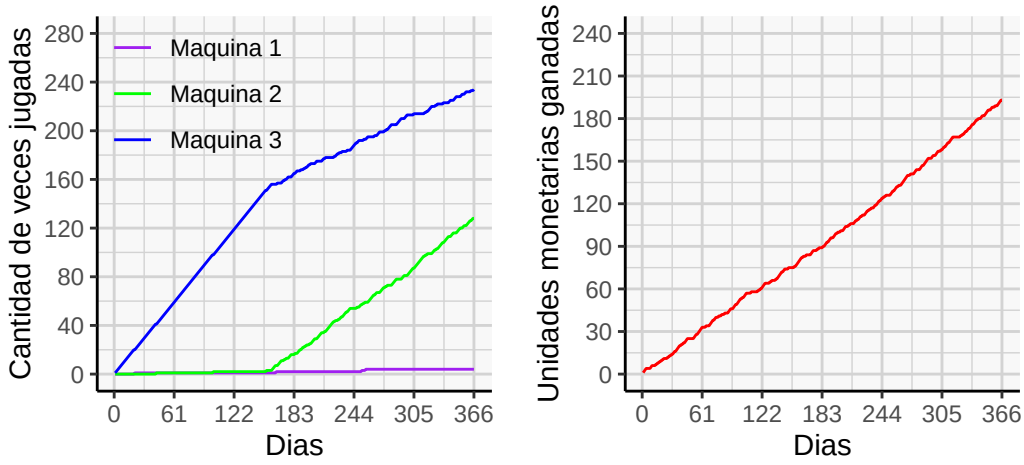
```
vector2 = data.frame(matrix(nrow = 50, ncol = 1))

temperatura = c(seq(1/25,2,1/25))

for (i in 1:50) {
  matriz2 = actualizar(366, 1000, "softmax", temp = temperatura[i] )
  matriz2 = filter(matriz2, Dia == 366)
  vector2[i,] = sum(matriz2[, 3:5])
}




xxx2 = as.matrix(vector2/1000)
xxx2 = as.data.frame(cbind(xxx2, temperatura))
colnames(xxx2) = c("x", "seq")


ggplot(xxx2)+geom_line(aes(x = seq, y = x), color = "red")+
tema+
scale_y_continuous(name="Ganancias promedio",breaks = seq(0,280,40),limits = c(0,280))+
  scale_x_continuous(name="Temperatura")+
  theme(plot.margin = margin(1.5, .3, 1.5, .3, "cm"))
```



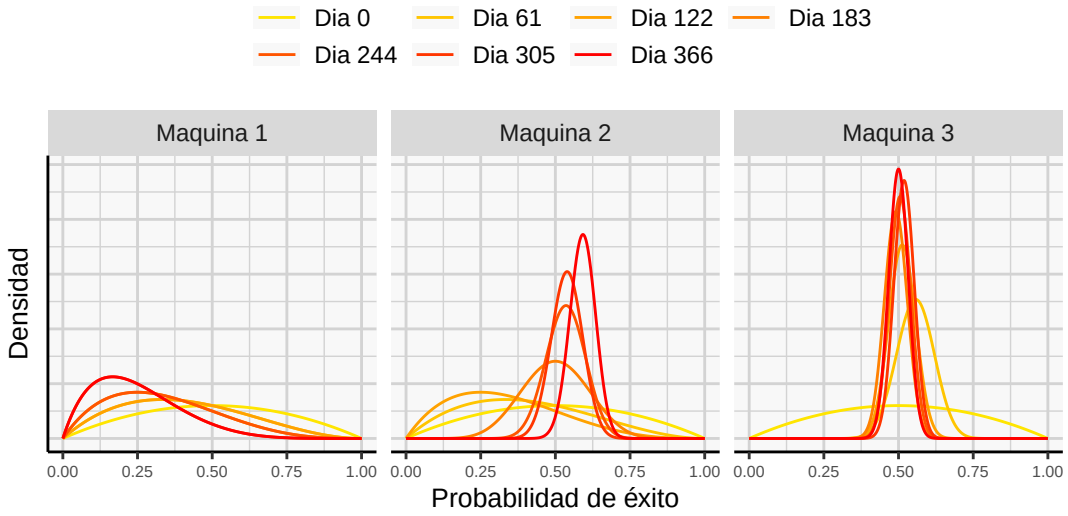

```
temp = xxx2$seq[which.max(xxx2$x)]
```

Queremos que en promedio se gane lo mayor posible. El valor que maximiza las ganancias es 0.12.

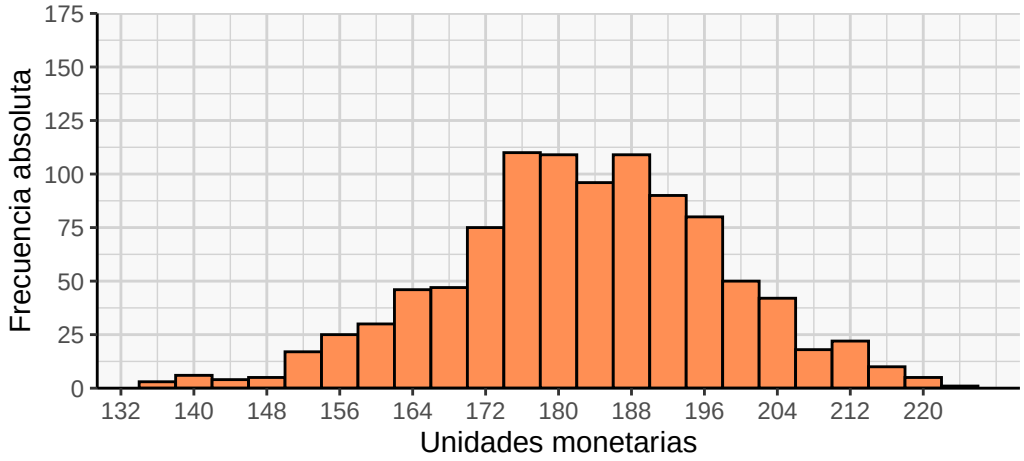
Se juega un año con esta estrategia y se obtienen los siguites resultados:



En los primeros cinco meses del año, se usa mayoritariamente la maquina 3 pero luego comienza a su vez a utilizar la maquina 2. Esto se debe a que el metodo permite la exploracion de las maquinas que no esta explotando. Al usar la maquina 2, aumenta su probabilidad de exito observada, que aumenta la probalidad de elegirla con el metodo. En el siguiente grafico se puede observar dicho efecto.



Para observar como varian las ganancias usando este metodo se simularon 1000 años, las cuales arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 183.22 unidades monetarias en un año.

Esto es un 9.29% menos que el caso ideal.

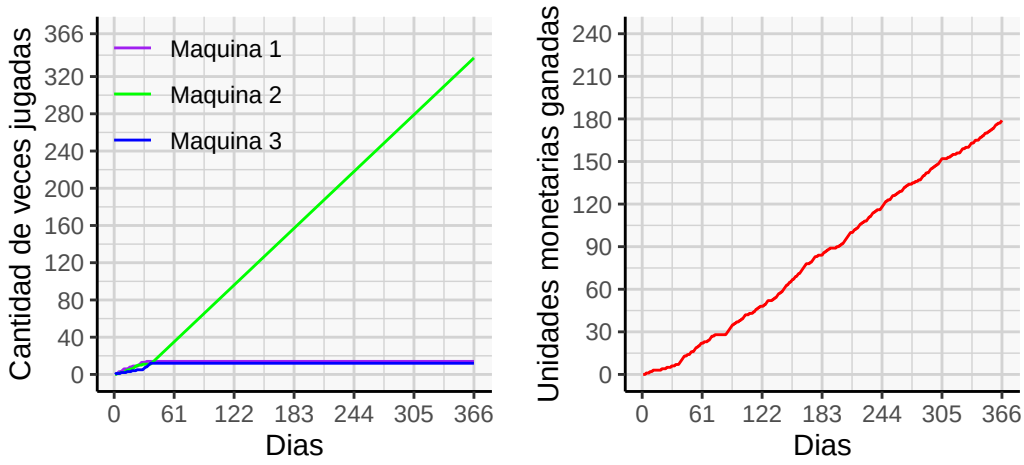
El 95% de las ganancias al final del año estuvieron entre 152 y 216. Su amplitud es de 60 unidades monetarias, siendo 1.71veces mas grande que el caso ideal.

Su ganancia promedio es menor que todas las estrategias greedy, pero la amplitud de su intervalo es mas baja. Es decir, si una persona quisiera tener mas seguridad sobre la cantidad de unidades monetarias a ganar jugaria esta a las mencionadas.

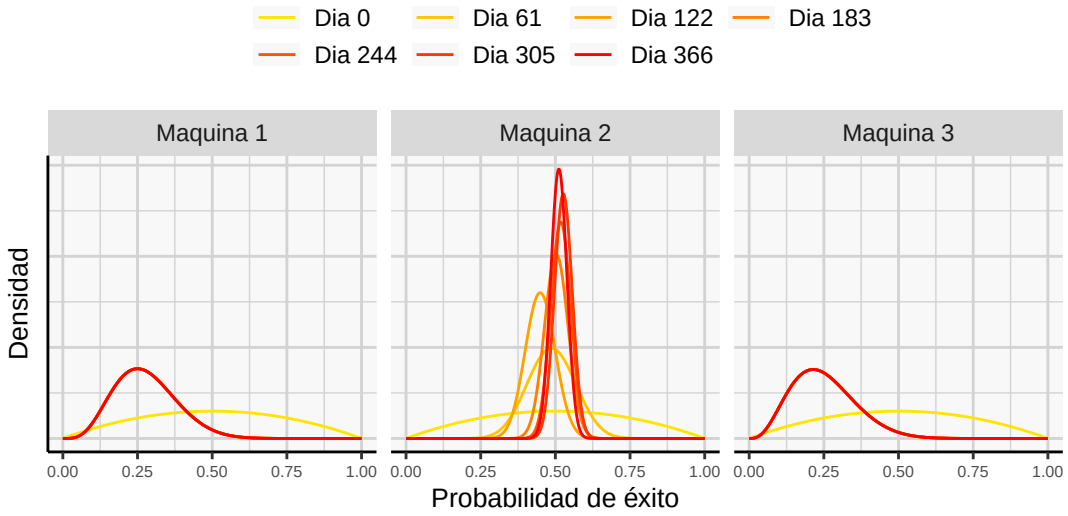
Estrategia 6: Upper-bound

Sera seleccionada la maquina que tenga el mayor extremo derecho de un intervalo de credibilidad. En este caso, el extremo derecho del intervalo usado es aquel que acumula el 97.5% de la densidad. Si los extremos derechos de los intervalos son iguales para algunas maquinas se elige una de ellas aleatoriamente.

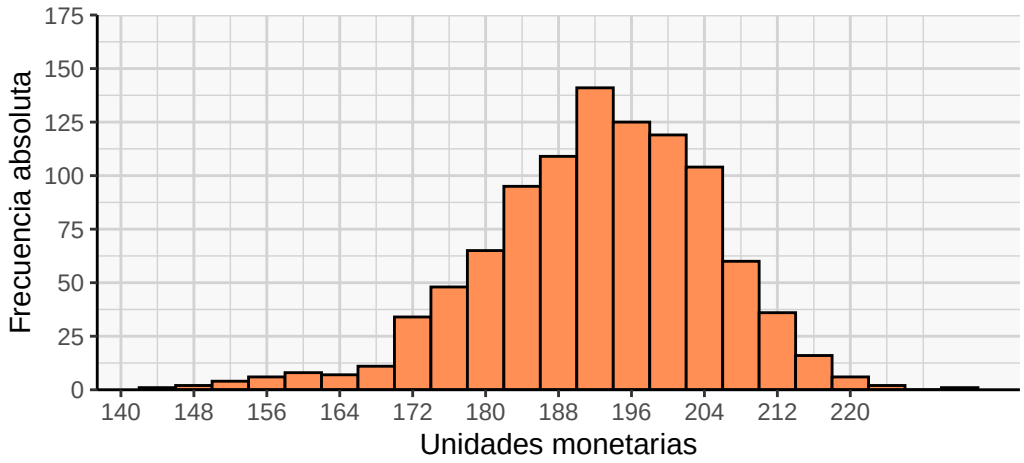
Se juega un año con esta estrategia y se obtienen los siguites resultados:



Al comienzo del año se juegan todas las maquinas en una proporcion parecida, pero al cabo de un mes y un par de dias el metodo elige utilizar exclusivamente la maquina 2.



Como varias de las estrategias anteriores, generalmente esta tambien luego de un cierto numero de dias procede a jugar en una sola maquina. Pero para poder observar su diferencia con las demas tecnicas se realizaron simulaciones de 1000 años, las cuales arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 192.9 unidades monetarias en un año. Esto es un 4.5% menos que el caso ideal.

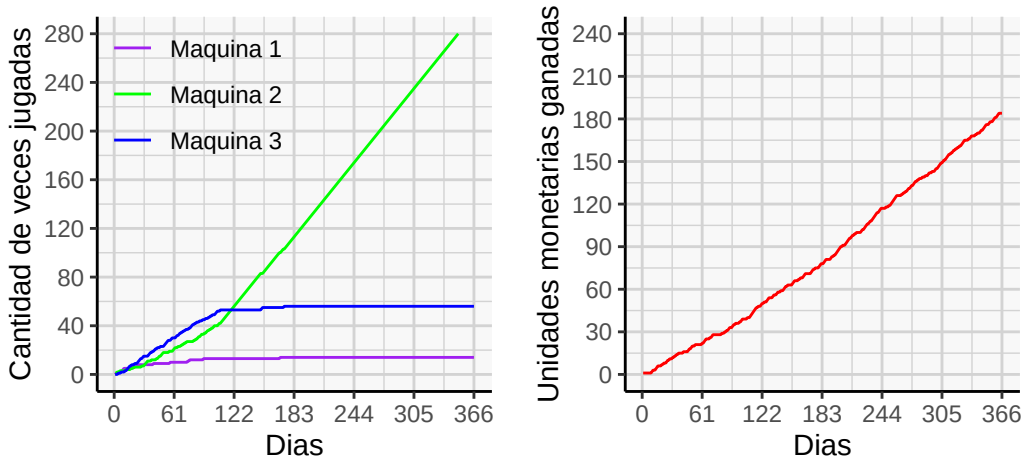
El 95% de las ganancias al final del año estuvieron entre 164 y 214. Su amplitud es de 50 unidades monetarias, siendo 1.43veces mas grande que el caso ideal.

Se puede observar que el histograma de esta estrategia es mucho mas simetrico que el de las tecnicas anteriores que se le parecen. Esto se debe a que cada año la maquina con la que mas veces se suele quedar el metodo es la 2, es decir, la que mayor probabilidad de exito tiene.

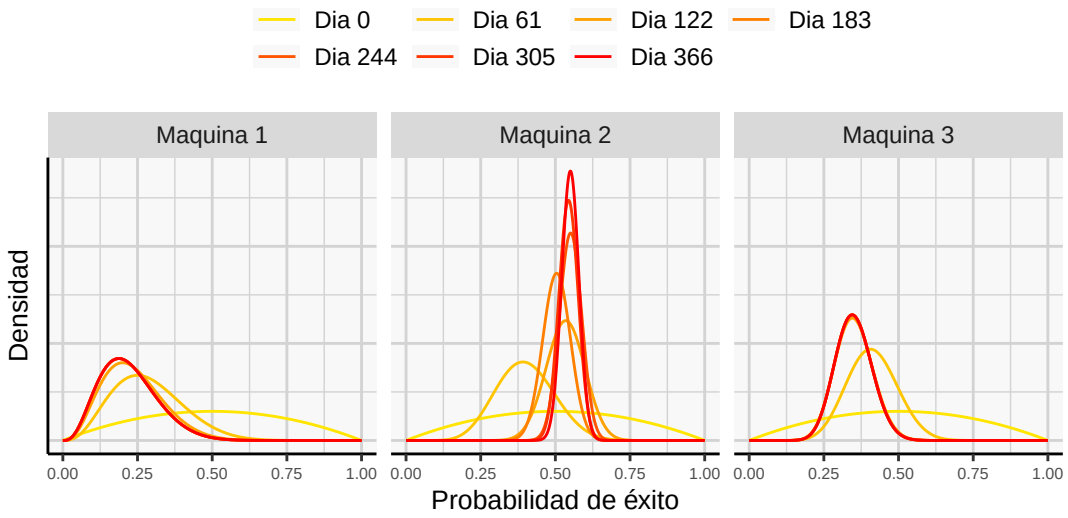
Estrategia 7: Thompson sampling

Para elegir que maquina se usa en el dia, se saca una muestra de la distribucion a porteriori de las probabilidades de éxito de cada una de ellas y se selecciona aquella con el maximo valor muestral.

Se juega un año con esta estrategia y se obtienen los siguites resultados:

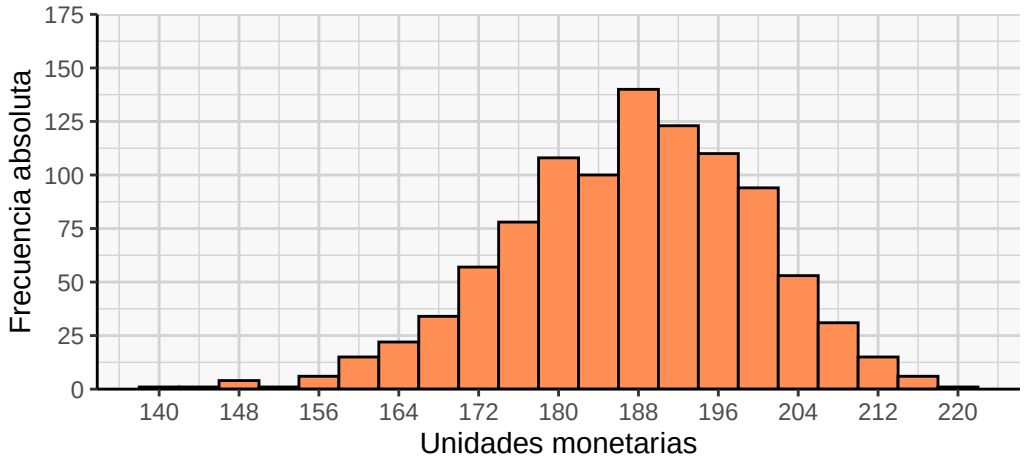


En este caso, en los primero 3 o 4 meses, la estrategia usa todas las maquinas. Luego de este periodo, utiliza mayormente la segunda dejando de utilizar las demas aproximadamente a mitad de año. Se parece a las demas estrategias con la diferencia de usar muchas mas veces las maquinas con las que no se termina quedando el metodo.



Generalmente esta estragia, como la mayoria de las demas, luego de un cierto numero de dias procede a jugar con una sola maquina.

Para observar como varian las ganancias usando este metodo se simularon 1000 años, las cuales arrojaron los siguientes resultados:



En promedio, siguiendo esta estrategia, se espera ganar 187.89 unidades monetarias en un año. Esto es un 6.98% menos que el caso ideal.

El 95% de las ganancias al final del año estuvieron entre 161 y 210. Su amplitud es de 49 unidades monetarias, siendo 1.4veces mas grande que el caso ideal.

Esta estrategia, como la anterior, tiene un histograma bastante simetrico. Es mejor que los otros metodos parecidos pero es peor que upper bound ya que tiene un promedio de ganancias mas chico, aunque la amplitud de su intervalo sea ligeramente mas pequeño.

Conclusion

En todas las estrategias usadas se vio entrelazado el concepto de explorar o explotar, dandole distintas importancias a cada una. Los mejores metodos son lo que pueden encontrar un equilibrio entre ambos enfoques. Aquel que explore lo justo y necesario para asegurarse de elegir la mejor opcion al momento de explotar. Viendo desde este enfoque los resultados, el mejor metodo entre los utilizados en el informe es el upper bound, siendo este el que mejor supo balancear ambos conceptos dando como consecuencia el mayor promedio de unidades monetarias ganadas y con casi la menor variabilidad entre todos las estrategias. Puede existir una mejor estrategia que capture mejor el concepto de “explorar vs explotar”.

Primera pagina que no tenga numero poner tildes decir si son metodos bayesianos anexo

Anexo