En BLS (Basic Lottery Sampling), para cada elemento distinto x_i con frecuencia absoluta f_i, se generan f_i *tokens* $t_{i,1}, \cdots, t_{i,f_i}$ que son variables independientes e idénticamente distribuídas con distribución uniforme en $(0,1)$. Para cada elemento se define su *ticket* $T_i = \max_{1\le j\le f_i}\{t_{i,j}\}$. Tenemos por tanto

$$\Pr\{t_{i,j} \le x\} = x, \qquad 1 \le i \le n, 1 \le j \le f_i \tag{1}$$

$$\Pr\{T_i \le x\} = x^{f_i}, \qquad 1 \le i \le n \tag{2}$$

Algunas propiedades claves de BLS son:

1. Si un elemento z está en la muestra S, entonces $t(z)$ es su ticket en curso, esto es, el mayor token generado para z hasta el momento.

2. Si un elemento z no está en la muestra S y el token t en curso para z le da "entrada" entonces t es el ticket de z, ningún token previo de z es mayor que t.

3. Si el número de elementos en S es m, entonces son áquellos con los m mayores tickets hasta el momento, es decir, los elementos que tienen asociados los tickets $T_{(1)}, \ldots, T_{(m)}$, donde $T_{(r)}$ es el r-ésimo mayor ticket en el conjunto $\{T_1, \ldots, T_n\}$.

Consideremos ahora la variante con "tickets paralelos" (h-LS). Para cada elemento z de la secuencia se genera un h-token, esto es, si z es la j-ésima ocurrencia del i-ésimo elemento diferente el h-token es un vector $\vec{t}_{i,j} = (t_{i,j}^{(1)}, \ldots, t_{i,j}^{(h)})$ de h tokens (variables uniformes en $(0,1)$ i.i.d.). Generalizando la definición anterior, podríamos decir que el h-ticket de x_i es $\vec{T_i} = (T_i^{(1)}, \ldots, T_i^{(h)})$ con $T_i^{(k)} = \max_{1\le j\le f_i}\{t_{i,j}^{(k)}\}$. El problema es que el criterio para que un elemento z entre o no en la muestra y la manera en que definimos el *threshold* para entrar en la muestra no generalizan BLS (y por tanto no tenemos propiedades equivalentes a las mencionadas para BLS).

Para empezar cuando z no esta en la muestra se genera el token $\vec{t}_{i,j}$ y se calcula el valor promedio

$$\bar{t}_{i,j} = \frac{1}{h} \sum_{1\le k\le h} t_{i,j}^{(k)},$$

que obviamente no está distribuido uniformemente en $(0,1)$. Si h es grande por la ley de los grandes números (teorema central del límite) $\bar{t}_{i,j}$ seguirá una distribución normal con $\mu = 1/2$ y varianza $\sigma^2 = 1/(12h) \to 0$.

Qué ocurre con el threshold? De cada elemento z en la muestra se almacena un vector $\vec{t}(z)$. Pero no es necesariamente el h-ticket de z ya que cuando un elemento se eyecta de la muestra dicho elemento no es el de h-ticket mínimo—no existe tal noción. Es decir, en BLS garantizamos que $t(z)$ es el máximo token generado para z hasta el momento, pero en la nueva variante

$$t(z)^{(k)} \neq \max_{1\le j\le f_i}\{t_{i,j}^{(k)}\}.$$

En cualquier caso de cada elemento en la muestra tendremos un valor promedio $\bar{t}(z)$ de $\vec{t}(z)$. Aunque las componentes no están distribuídas uniformemente ni tampoco son máximos de f_i uniformes siguen siendo i.i.d. y por lo tanto la distribución de $\bar{t}(z)$ tiende hacia una distribución normal a medida que h aumenta. Aunque $t(z)^{(k)}$ no se distribuye como el máximo de f_i uniformes independientes, se acercará más o menos a ello, y por tanto para $\bar{t}(z)$ tendriamos $\mu \approx f_i/(f_i+1)$ y $\sigma^2 \to 0$—el valor esperado del máximo de f_i uniformes en $(0,1)$ es $f_i/(f_i+1)$.

Por último, el threshold se define como el mínimo de los $\bar{t}(z)$, es decir, el m-ésimo de un conjunto de las m variables aleatorias independientes pero no idénticamente distribuídas. Esto también es cierto en LS, pero el conjunto de las m variables son máximos de f_i variables uniformes, mientras que en h-LS son "normales" con $\mu \approx f_i/(f_i+1)$.

En conclusión, h-LS **no** es equivalente a h ejecuciones en paralelo de LS. En este último caso, cada ejecución de LS acabaría con una muestra distinta S_k, $1 \leq k \leq h$ y no parece haber una relación simple entre la muestra S que genera h-LS y las h muestras S_k.

Todo ello no obstante, el problema más serio para h-LS es que para decidir si un elemento z se incorpora o no a la muestra se use $\bar{t}_{i,j}$ que, como hemos visto, tendrá valor promedio $1/2$ (y a medida que aumente h mayor será la concentración entorno al valor medio). Conviene pues usar una uniforme para decidir si se entra o no (digamos $t_{i,j}^{(k)}$ para alguna k), pues aunque el valor medio seguirá siendo $1/2$ habrá suficiente varianza para que la probabilidad de que el elemento se incorpore a la muestra no sea increiblemente pequeña.

Si en h-LS usamos el promedio $\bar{t}_{i,j}$ del h-token $\vec{t}_{i,j}$ para decidir si un elemento entra o no, para un elemento con frecuencia f_i el valor esperado del máximo de los $\bar{t}_{i,j}$ será, aproximadamente,

$$\frac{1}{2} + \sqrt{1}12h\Phi^{-1}\left(\frac{f_i}{f_i+1}\right),$$

donde $\Phi(x)$ es la cdf de una normal con $\mu = 0$ y $\sigma^2 = 1$. Esta aproximación no es muy buena, pero indica lo mismo que la intuición: si h es muy grande todos los f_i pseudo-tokens $\bar{t}_{i,j}$ serán muy cercanos a $1/2$ y será casi imposible entrar en la muestra (salvo al principio, de manera que la muestra quedará muy rápidamente "fijada"). O quizás sea al revés: el threshold se distancia muy, muy lentamente de $1/2$ y hay mucha "volatilidad" en los contenidos de la muestra, reflejando en muy pequeña medida las diferentes frecuencias de los elementos.

El análisis de h-LS se torna complicado; lo que sí parece claro es que la decisión de entrar o no en la muestra no puede hacerse en base a los pseudo-tokens $\bar{t}_{i,j}$. También apunta a que un valor moderado de h tendrá el efecto beneficioso de eliminar los *outliers* de LS (elementos infrecuentes con tickets demasiado altos); pero un valor demasiado alto de h suaviza tanto las diferencias de frecuencia que nos lleva a una situación indeseable. Cuantificarlo apropiadamente resulta muy complicado, pero de todos modos estos razonamientos aproximados nos ayudan a entender los fenomenos observados en los experimentos.