Proyecto 1: Tópicos en Manejo de Grandes Volúmenes de Datos

Sketches: Estimadores de frecuencia

Docente: Cecilia Hernández

Alumno: Juan Albornoz

3 de Noviembre del 2020

1. Aspectos relevantes en la implementación de los algoritmos:

1.1. CountMin y CountMinCU:

- Utilicé substr() para el parsing de los q-mer en las secuencias genómicas. Para un rango de 0 a un límite definido como (longitud de la línea – longitud del q-mer) me fue posible identificar cada uno de estos q-mer.
- Función hash utilizada: **MurmurHash3**
    - **MurmurHash3_x86_32**(qmer, (uint16_t)strlen(qmer), seed, hash_otpt)
    - **Parámetro de tipo entero (seed):** Le brinda a la función la capacidad de ser "randomizada", esto es, entregar valores hash distintos para valores de entrada iguales. Consideré este parámetro en un rango de 0 a d, ayudándome a definir la d_i función hash dentro de mis funciones hash a utilizar.
    - La función retorna valores enteros sin signo en el rango del uint32_t, esto es un rango de 0 a 2^{32} valores posibles, pero que es posible acotar mediante casting a valores en el rango del uint16_t, esto es 2^{16} posibles valores.

2. Evaluación experimental:

Características de hardware y sistema sobre el que se realizó la evaluación:

- Intel® Core(TM) i5-8250U CPU @ 1.60GHz-1.80GHz
- Memoria RAM: 8.00GB (7,86GB utilizable)
- Sistema operativo de 64 bits, procesador x64, Windows 10 Home Single.

Los archivos de muestra que utilice fueron los siguientes:

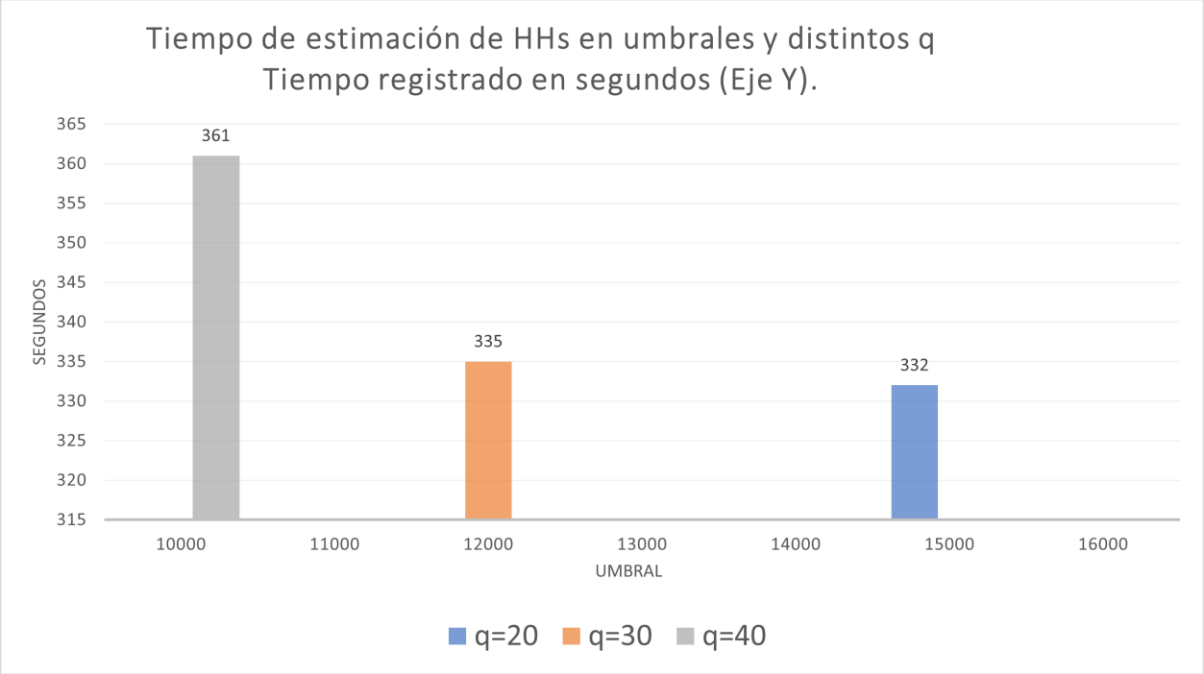
- **ERR626208.fastq** (Muestra 1)
- **ERR626210.fastq** (Muestra 2)

En primera instancia presento gráficos que indican el tiempo de estimación de los Heavy Hitters para distintos valores q {20, 30, 40}, y en umbrales escogidos.

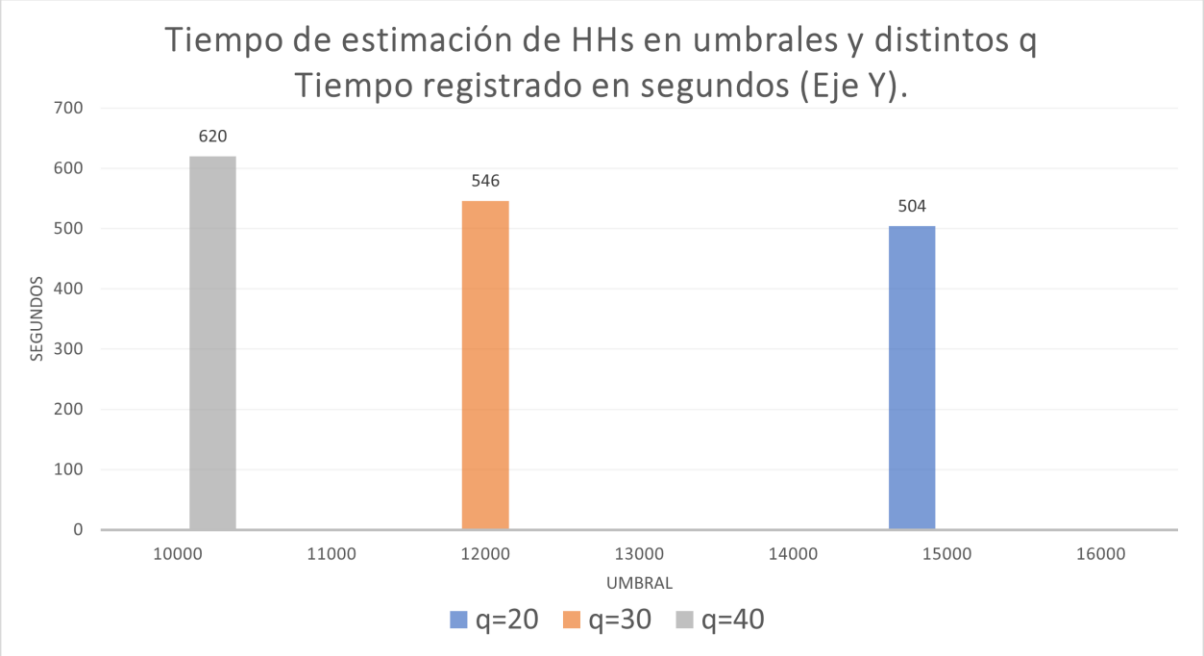
2.1. Gráficas del tiempo de estimación versus los valores q en distintos umbrales:

*Hay que destacar que solo incluí el tiempo de procesamiento para las ejecuciones en el que el umbral definido era el adecuado para el cálculo de los Heavy Hitters.

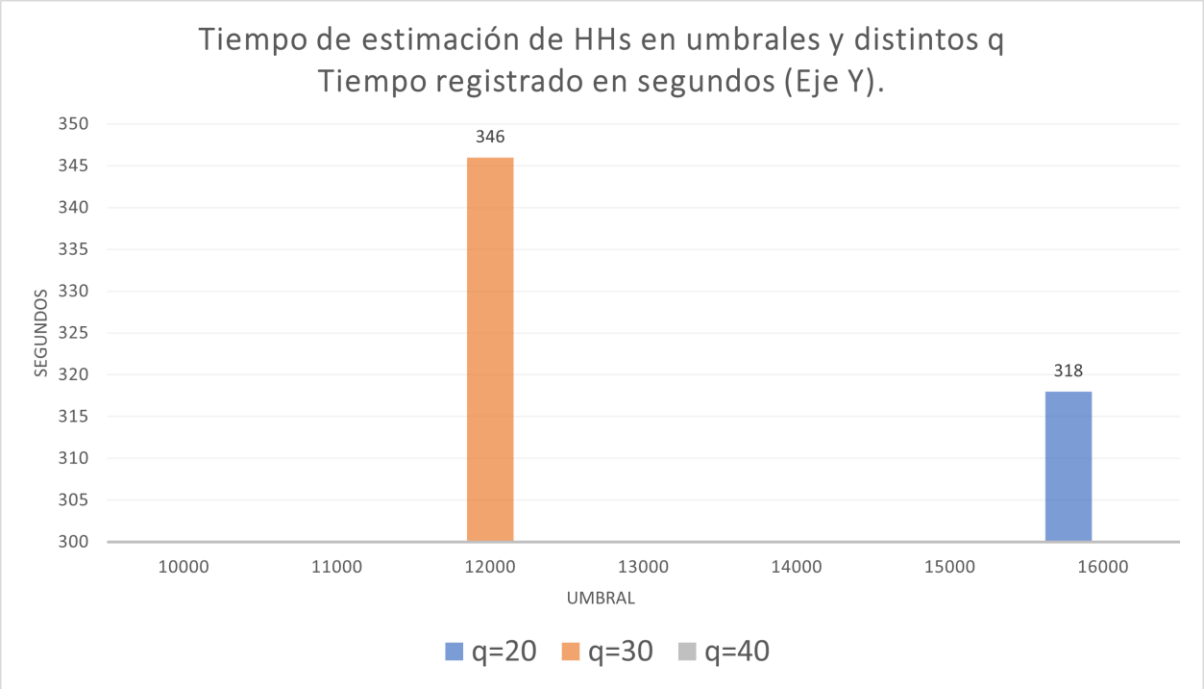
- CountMin, Parámetros d=4, w=2^{16}: (Muestra 1)



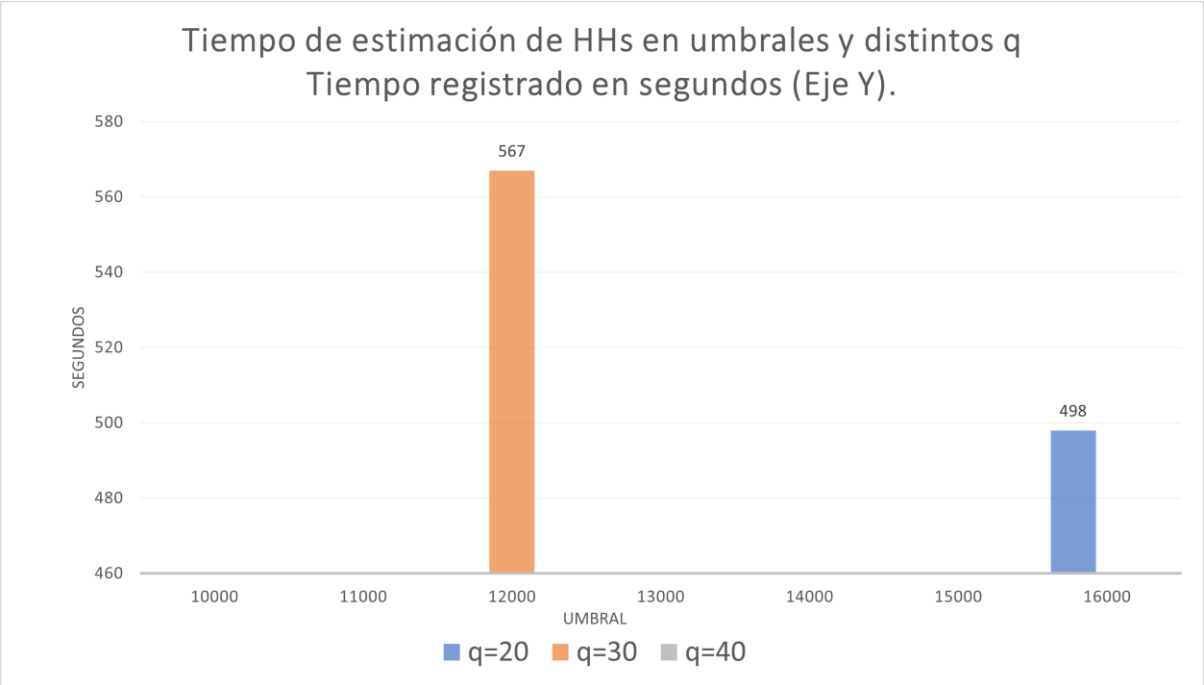
- CountMin, Parámetros d=8, w=2^{16}: (Muestra 1)



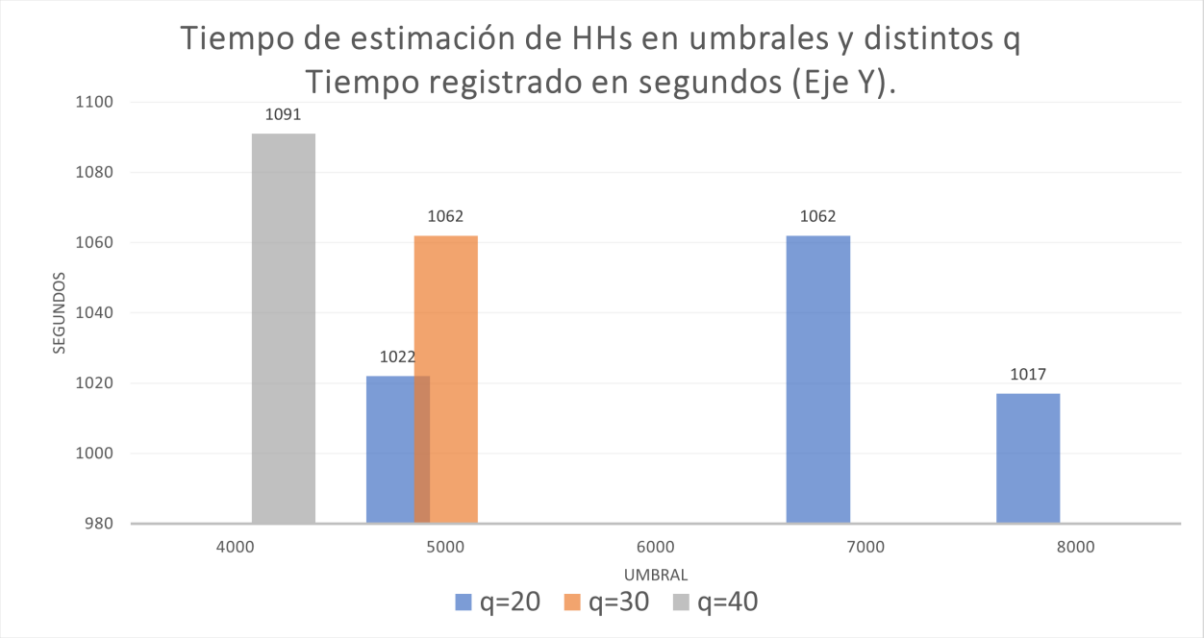
- CountMin, Parámetros d=4, w=2^{16}: (Muestra 2)



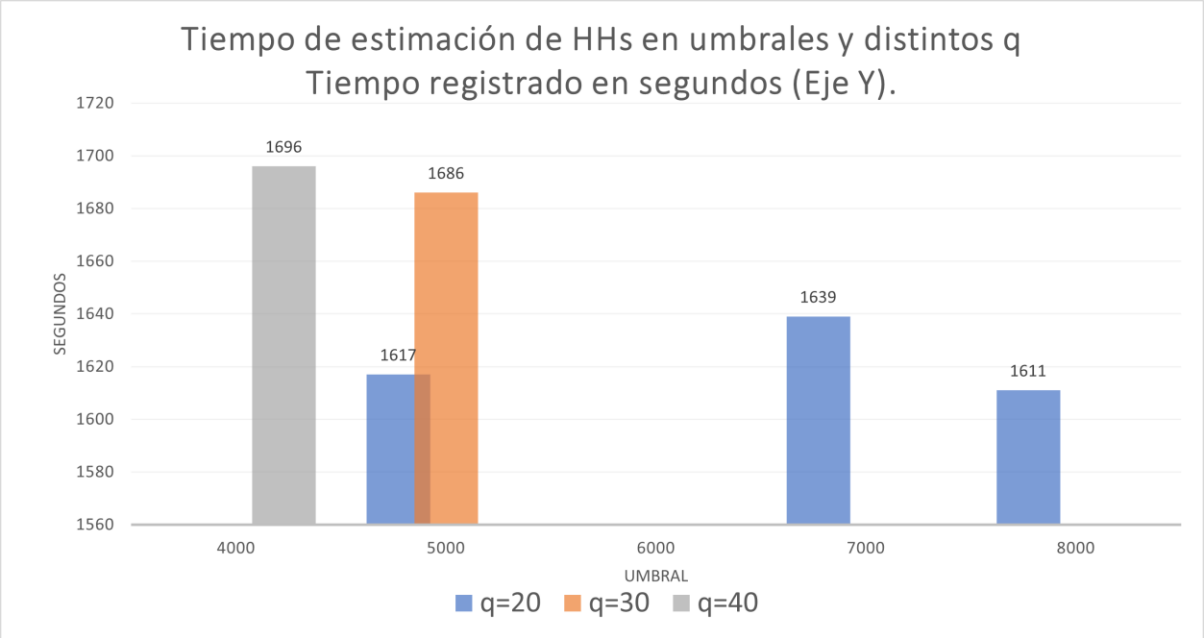
- CountMin, Parámetros d=8, w=2^{16}: (Muestra 2)



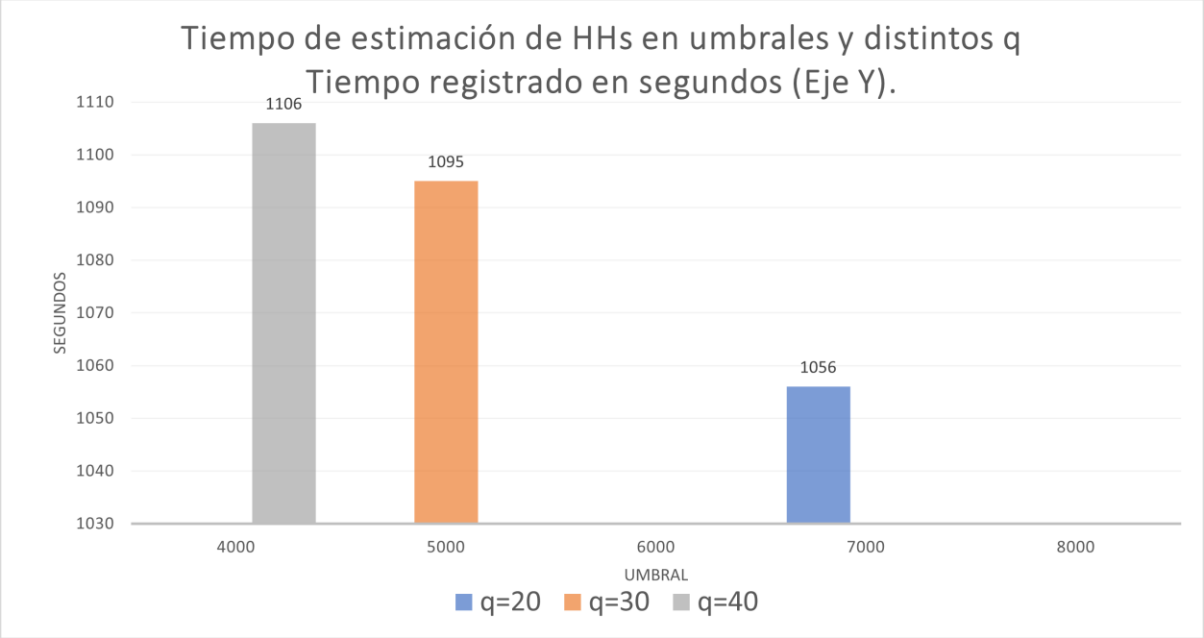
- CountMinCU, Parámetros d=4, w=2^{16}: (Muestra 1)



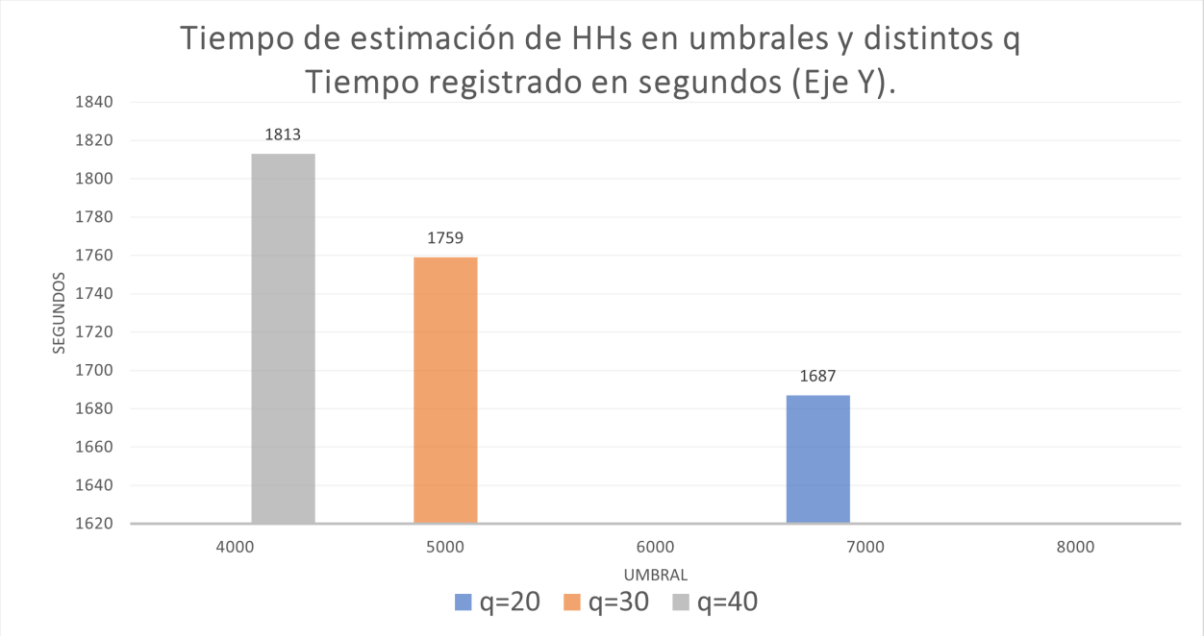
- CountMinCU, Parámetros d=8, w=2^{16}: (Muestra 1)



- CountMinCU, Parámetros d=4, w=2^{16}: (Muestra 2)



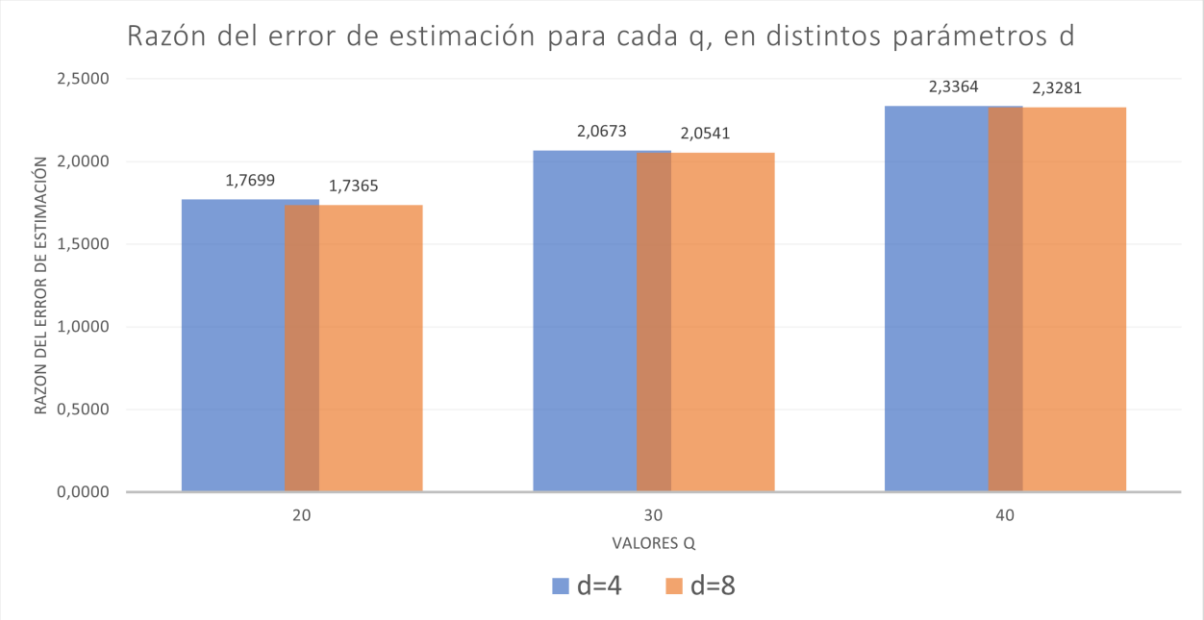
- CountMinCU, Parámetros d=8, w=2^{16}: (Muestra 2)



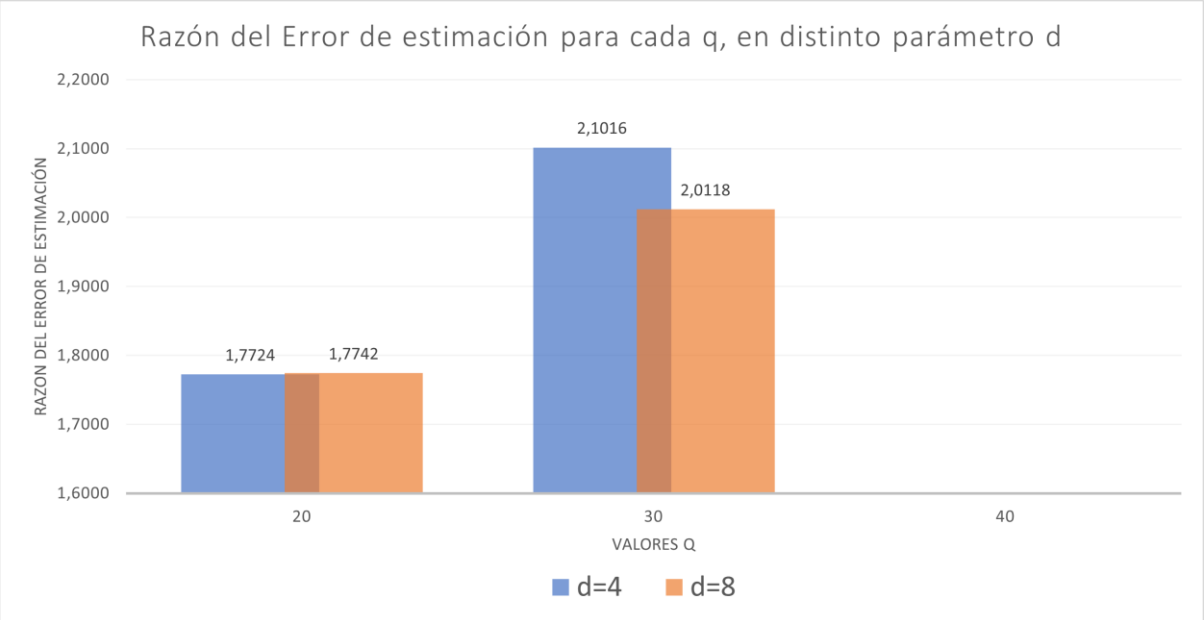
2.2. Gráficas de la razón del error de estimación (Rs):

*Calcule cada razón mediante la raíz n-ésima de la productoria de la razón de cada frecuencia estimada con respecto a su frecuencia real.

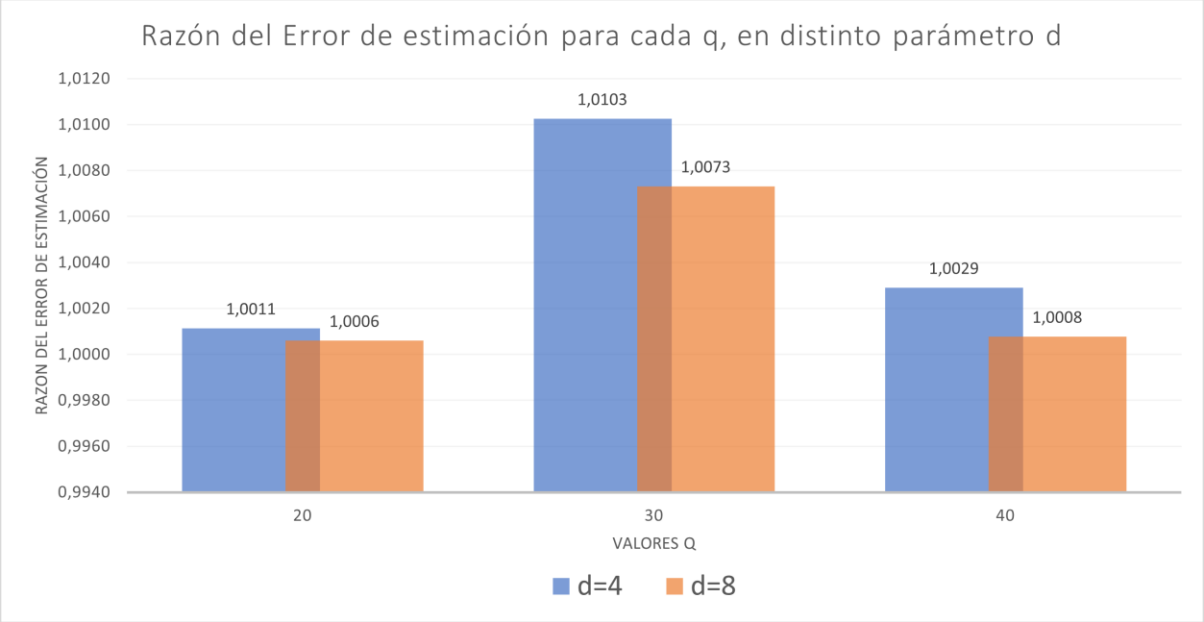
- CountMin, Parámetros d=4 y d=8, w=2^{16}: (Muestra 1)



- CountMin, Parámetros d=4 y d=8, w=2^{16}: (Muestra 2)



- CountMinCU, Parámetros d=4 y d=8, w=2^{16}: (Muestra 1)



- CountMinCU, Parámetros d=4 y d=8, w=2^{16}: (Muestra 2)