Tarea 3
Distribuciones Discretas

Erick Cervantes Mendieta
Matrícula: 2032430

Modelos Probabilistas Aplicados

Septiembre 2020

1. Presentación de los datos

Para esta tarea se analizó el libro titulado "The Singing Mouse Stories", el cual se encuentra disponible de manera gratuita en el sitio Web: *Project Gutenberg* [1]. Se hizo un análisis en el programa R (Versión 4.0.2) [2] de la frecuencia en que aparece la letra *e* en todo el libro, por ejemplo, en el título del libro podemos observar que dicha letra aparece en el lugar número tres, luego se encuentran otras doce letras para aparecer de nuevo y finalmente se anotaron otras cinco letras para volver a tipear *e*, un análisis parecido se hizo con la palabra *singing*. Finalmente se presenta un análisis del número de veces en que aparece la frase *the singing mouse* en el texto.

2. Análisis de los datos

La figura 1(a) muestra el comportamiento de las veces en que fue apareciendo la letra e, en dónde, a simple vista se podría decir que sigue una distribución Geométrica, sin embargo la primera barra podría causar un poco de confunsión, por lo que en la figura 1(b), 1(c) y 1(d) se muestran algunas simulaciones de las distribuciones Geométrica, Binomial y Binomial Negativa, respectivamente, utilizando algunos parámetros del libro utilizado.

Se puede concluir que al parecer la frecuencia con la que fue utilizada la letra *e*, sigue una distribución Geométrica, aunque hace falta hacer un análisis más profundo para poder comprobar este supuesto.

En la figura 2(a) se puede visualizar el número de veces en la que fue apareciendo la palabra *singing*, se puede observar que al parecer no sigue alguna distribución de probabilidad discreta, ya que tiene un gran pico casi al inicio de la gráfica, quizá al suavizar un poco la gráfica se podría observar algo más. Luego, se trato de simular como si fuera una Binomial,

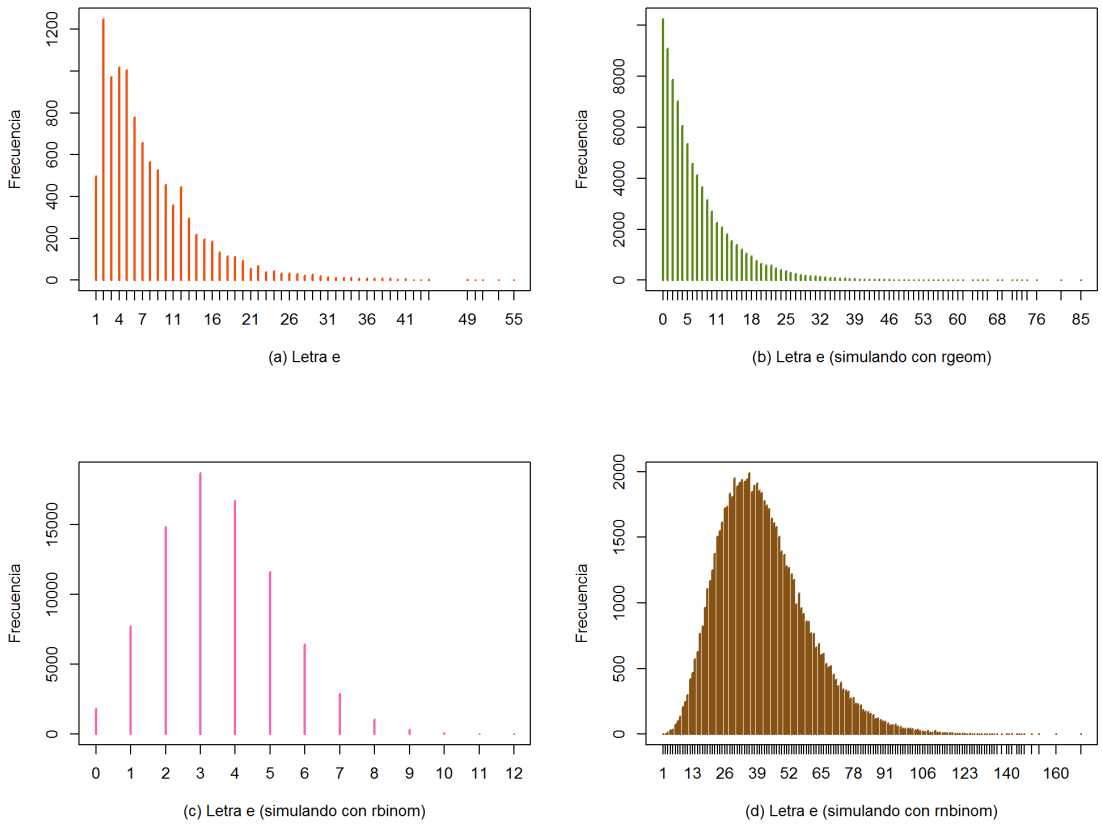

Figura 1: Frecuencia de la letra *e* utilizadas en el texto

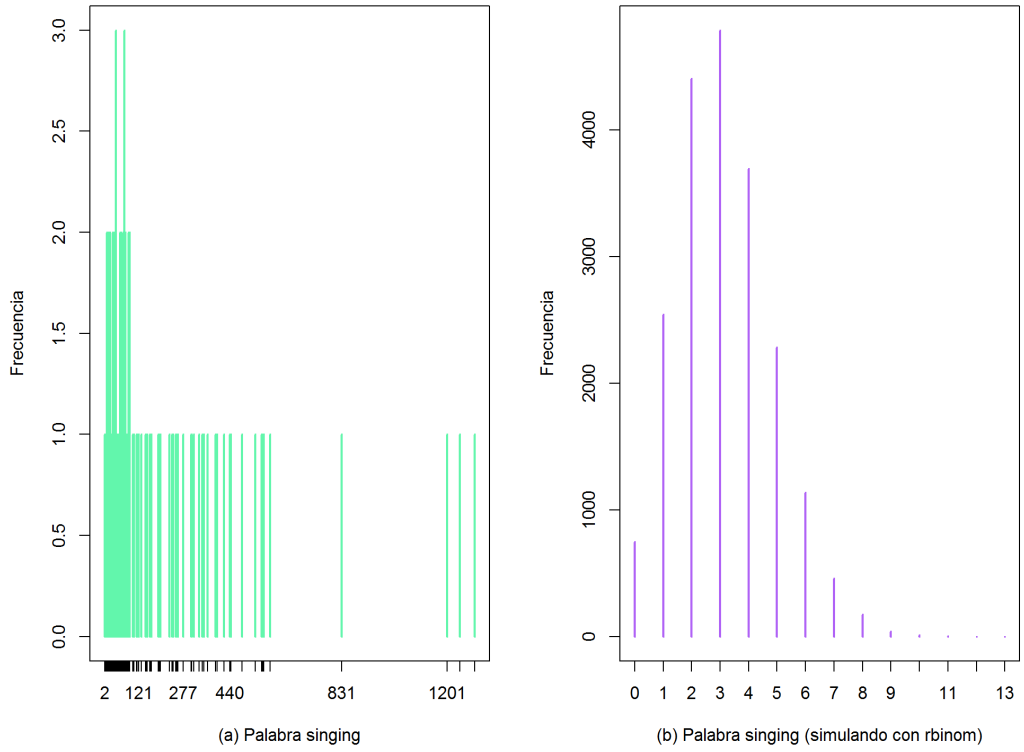

Figura 2: Frecuencia de la palabra *singing* en el texto

sin embargo no se pudo concluir visualmente que se siga dicha distribución.

En la figura 3 se puede apreciar la frecuencia con la que fue apareciendo la frase *the singing mouse*, la cual se escogió por hacer referencia al título del libro, por lo que se suponía tendría que aparecer bastantes veces en el texto como lo muestra el lado izquierdo de la figura, sin embargo, existen grandes huecos para que vuelva a aparecer, lo que hace suponer que no siga alguna distribución de probabilidad.

Finalmente, en la figura 4 se aprecia el comportamiento de la longitud de los trigramas de todo el texto, en donde, se puede apreciar que siguen una distribución Binomial Negativa, sería interesante ver el comportamiento de los demás ngramas, y ver si se puede concluir con alguna métrica que pudo haber seguido el autor.

Referencias

[1] Emerson Hough. *The Singing Mouse Stories*. Library of Alexandria, 1895.

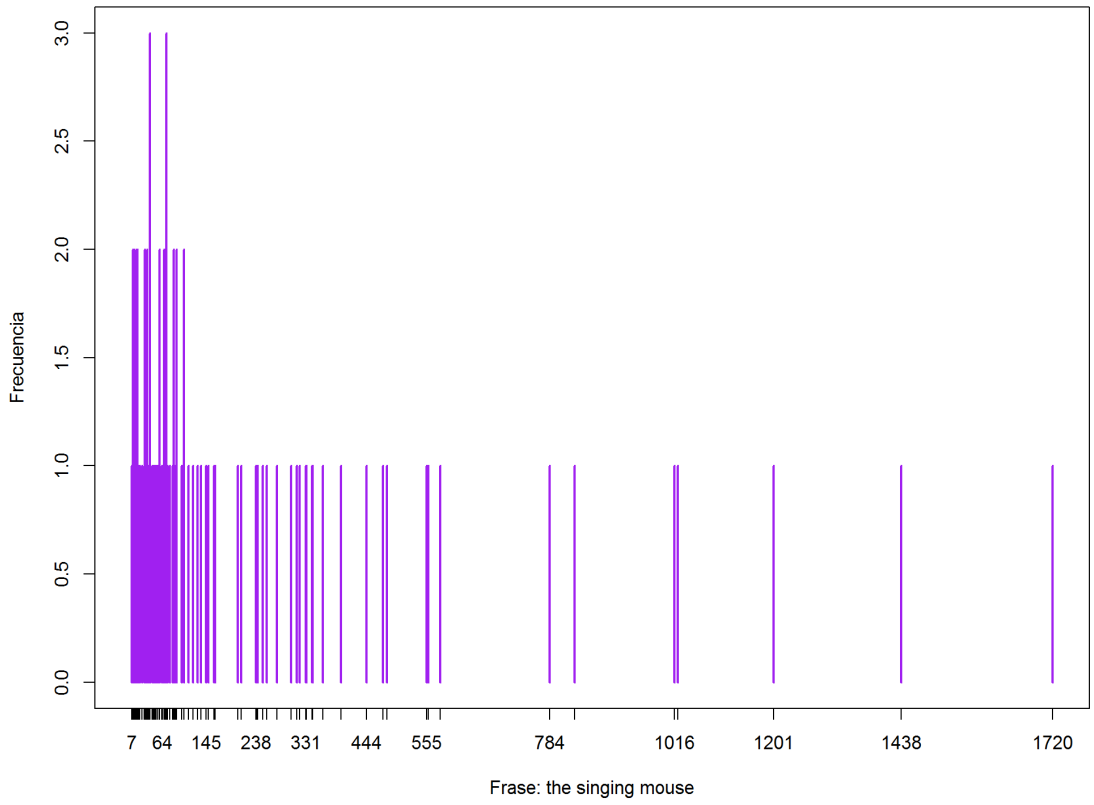

Figura 3: Frecuencia de la frase *the singing mouse* en el texto

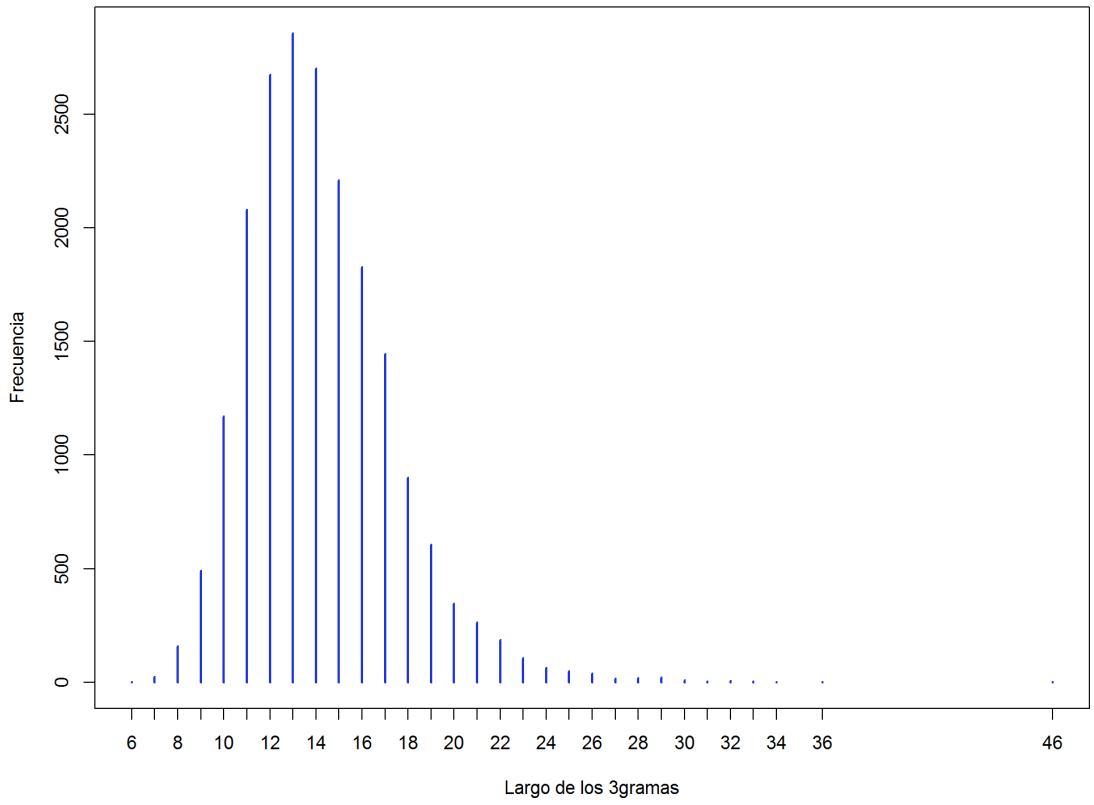

Figura 4: Frecuencia del largo de los gramas con $n = 3$

[2] R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/, 2020.