Tarea 3 de Modelos Probabilistas Aplicados

Distribuciones de Probabilidad

5271

22 de septiembre de 2020

1. Origen de los datos

En este trabajo se utiliza como fuente de los datos el libro "The Adventures of Sherlock Holmes"[1], del escritor y médico británico Arthur Conan Doyle. Este libro se encuentra disponible en la biblioteca virtual gratuita Project Gutenberg, con el siguiente enlace: https://www.gutenberg.org.

2. Sobre el libro

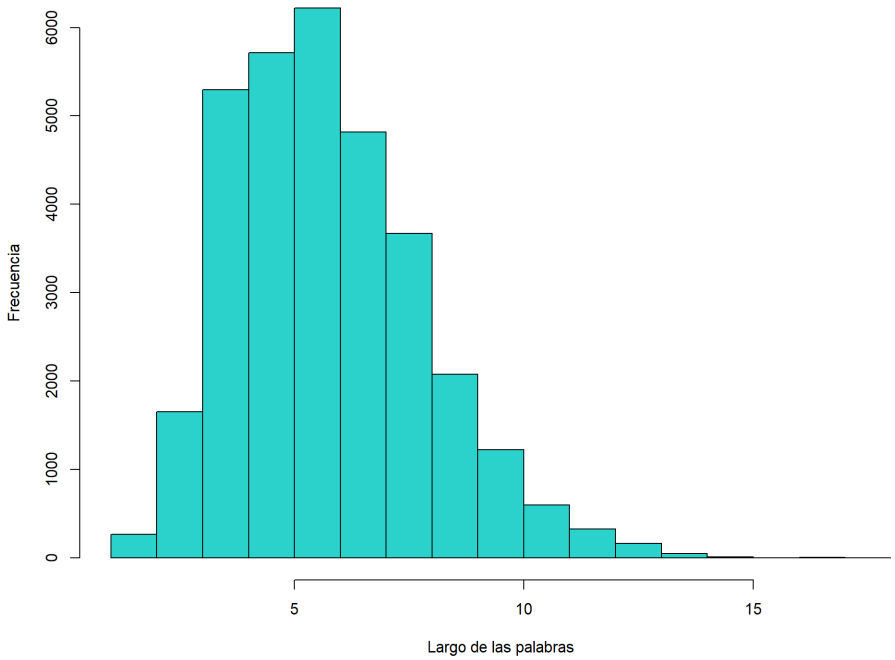
Este libro es una colección de doce cuentos de Arthur Conan Doyle, los cuales fueron publicados por primera vez el 14 de octubre de 1892. El mismo agrupa los primeros cuentos con el detective consultor Sherlock Holmes, que se habían publicado en doce números mensuales de The Strand Magazine de Julio de 1891 a junio de 1892.

3. Análisis y tratamiento de los datos

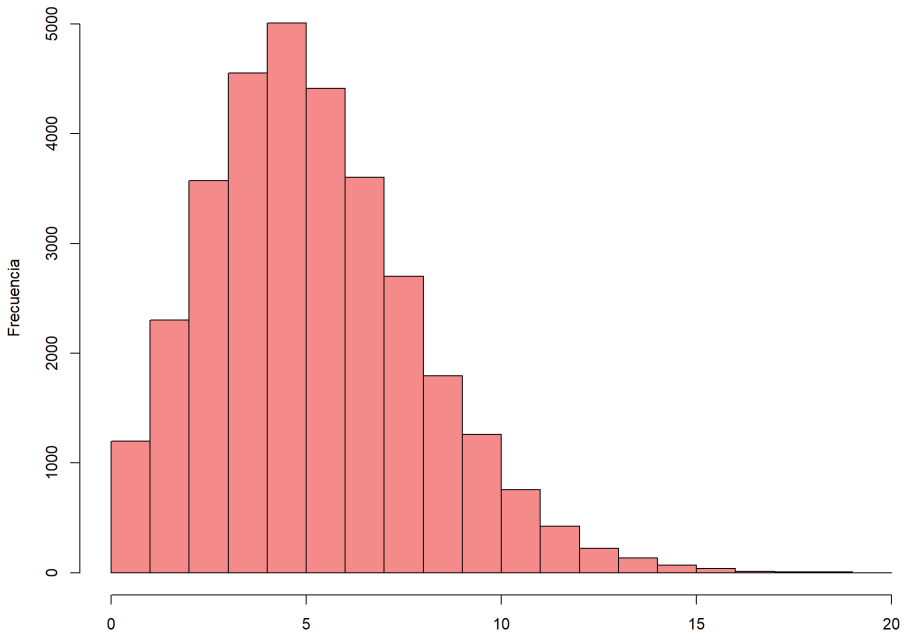
Al libro seleccionado se le realiza un estudio de frecuencia de ocurrencia de la cantidad de letras por palabras, de palabras por oraciones y cantidad de oraciones por párrafo. El análisis será realizado en el programa R versión 4.0.2 [2] en el entorno de desarrollo Rstudio [3].

3.1. Letras por oraciones

De le libro en cuestión se extraen todas las palabras que componen el texto, de las mismas se eliminan la llamadas palabras "vacías", es decir las que no aportan contenido al libro, el resultado del filtro se muestra en el cuadro 1 de la página 3. Una vez realizado este filtrado se procede a contar la cantidad de letras que tiene cada palabra, como se observa en la figura 1(a) de la página 5. En la figura 1(b) de la página 5, se muestra la distribución binomial negativa con parámetros $n =$ cantidad de palabras después del filtro, $k = 18$ y $p = 0{,}75$, a la que se asemeja la distribución de las cantidad de letras por palabras en el texto analizado.



(a) Cantidad de letras por palabras



(b) distribución binomial negativa creada con la función *rnbinom* de R

Figura 1: Histogramas de distribución de cuantiad de letras por palabras en el texto

Cuadro 1: fragmento de *Data frame* resultante del filtrado

	Gutenberg_id	Palabras
2	1661	sherlock
3	1661	holmes
19	1661	mystery
20	1661	orange
167	1661	crime
168	1661	occupied
169	1661	immense
170	1661	faculties
171	1661	extraordinary
172	1661	powers

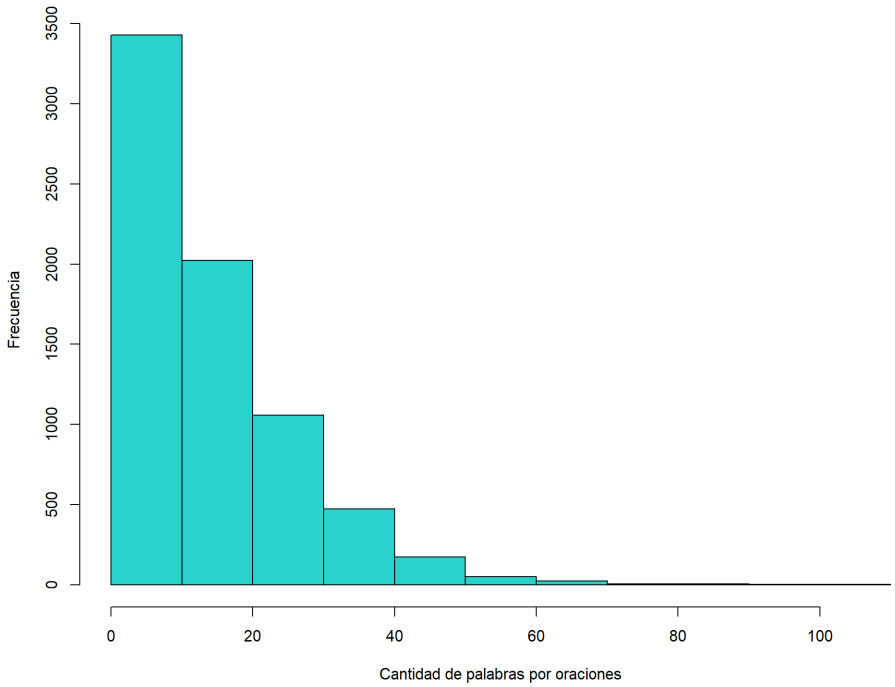
3.2. Palabras por oraciones

Para el trabajo con las oraciones, se extraen todas las presentes en el texto y se cuenta la cantidad de palabras que conforman cada una de estas oraciones. Lo anterior se puede observar en la en la figura 2(a) de la página 4. En la figura 2(b) de la página 4 se muestra una distribución geométrica con parámetros n = cantidad de oraciones del texto y $p = 0{,}055$. Dicha distribución se asemeja a la distribución de la cantidad de palabras en las oraciones del texto.

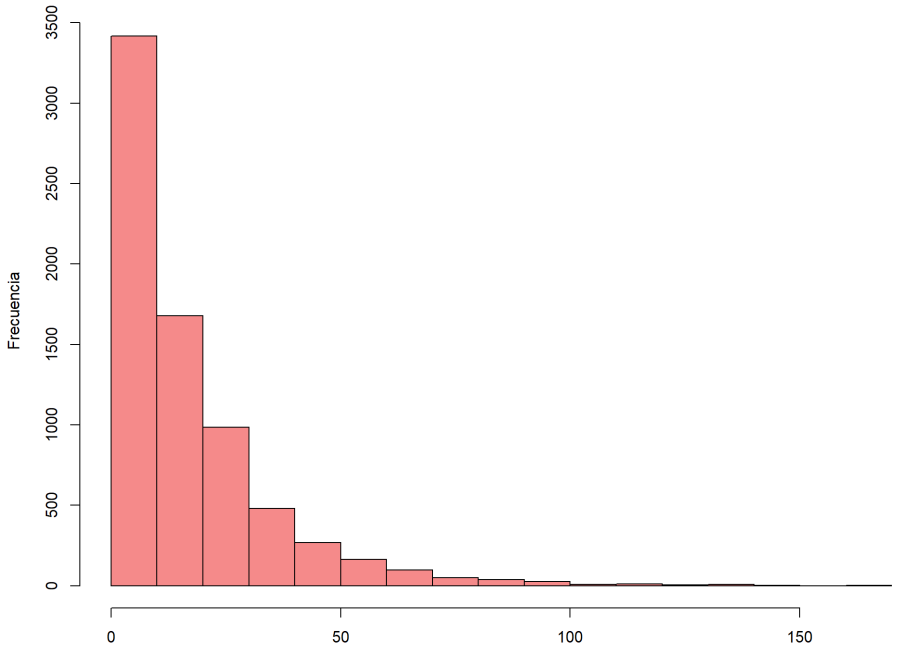
3.3. Oraciones por párrafos

Análogamente a lo realizado en la subsecciones anteriores se procede al análisis de la cantidad de oraciones por párrafo, como se puede observar en la figura 3(a) de la página 5. En la figura 3(b) de la página 5 se muestra una distribución geométrica con parámetros n = cantidad de párrafos del texto y $p = 0{,}29$. La distribución mostrada en figura 3(b) es muy similar a la distribución de la cantidad de oraciones por párrafos en el libro.

El código general se encuentra disponible en el repositorio https://github.com/Albertomnoa/Tareas

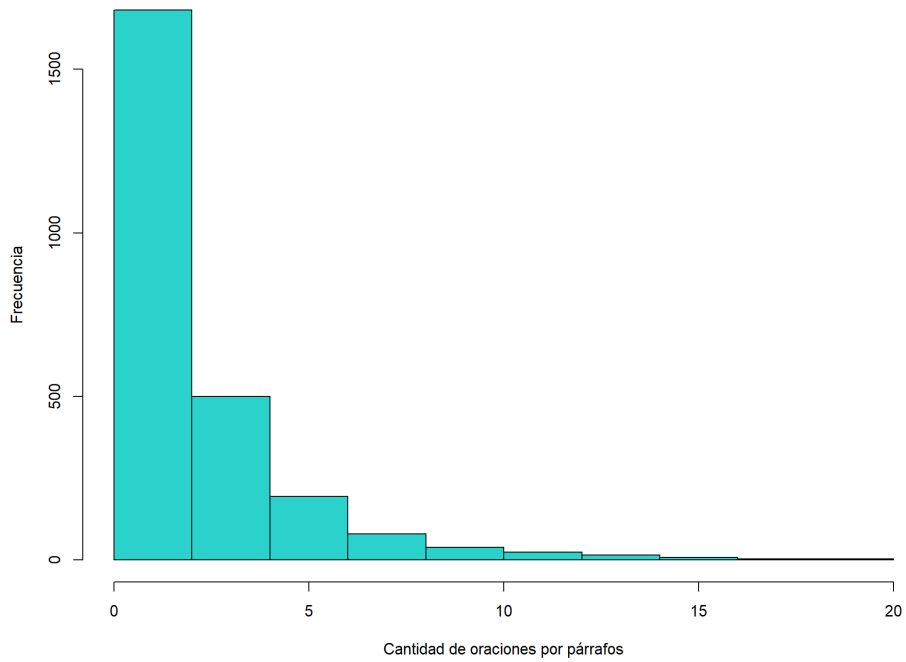


(a) Cantidad de palabras por oraciones

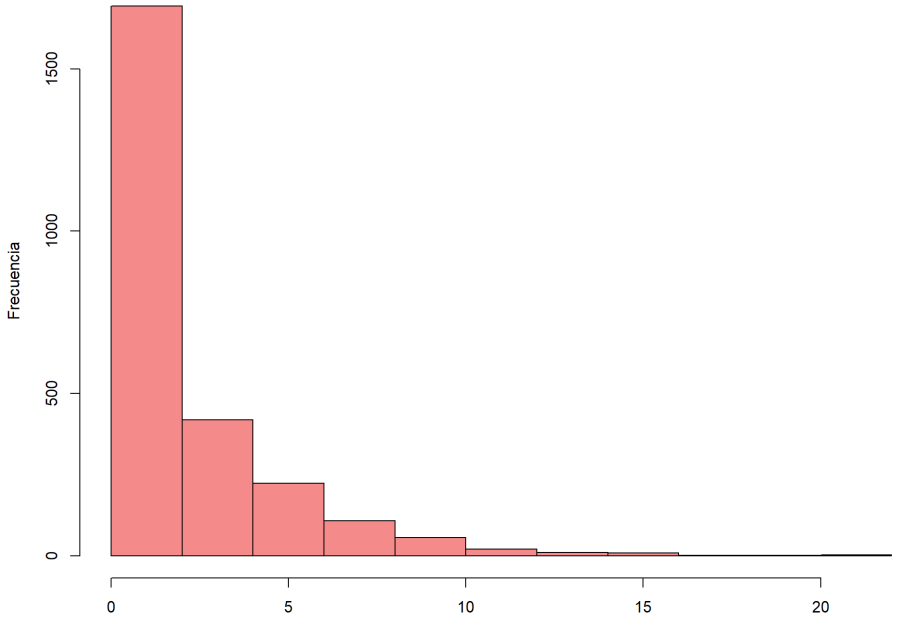


(b) distribución geométrica creada con la función *rgeom* de R

Figura 2: Histogramas de distribución de cuantiad de palabras por oraciones en el texto



(a) Cantidad de oraciones por párrafos



(b) distribución geométrica creada con la función *rgeom* de R

Figura 3: Histogramas de distribución de cuantiad de oraciones por párrafos en el libro

Referencias

[1] Arthur Conan Doyle. *The Adventures of Sherlock Holmes*. George Newnes, United Kingdom, 1892.

[2] R Core Team. R: Un lenguaje y un entorno para la informática estadística, 2020.

[3] RStudio Team. Rstudio: Entorno de desarrollo integrado para R, 2020.