Distribuciones de probabilidad geométrica, binomial y binomial negativa.

Gabriela Sánchez Y.
5064

En el presente trabajo se analiza el tipo de distribuciones que pueden están presentes en el texto del libro *"Anne of Green Gables"* [4] que puede obtenerse de manera gratuita en el sitio de Project Gutenberg.

1. Introducción

El procesamiento del texto se realiza con la ayuda del lenguaje de programación R versión 4.0.2 [2], haciendo uso de distintas librerías:: `gutenbergr` que permite acceder al texto plano del libro y, `tidytext` y `dplyr` que permiten la descomposición del texto.

En el preprocesamiento se eliminan las líneas vacías, el índice y los guiones bajos cuando hay palabras con guiones bajos alrededor para indicar énfasis [6]. Para tener un punto de comparación sobre lo que es propio del autor en la escritura de un texto se elige un cuento cuya fecha de publicación no se aleja mucho de la del libro base. En este caso se utiliza el libro *Peter Pan* [1] que se descarga del mismo sitio [3].

El procedimiento realizado para el análisis puede encontrarse en el código `t3.R` [7].

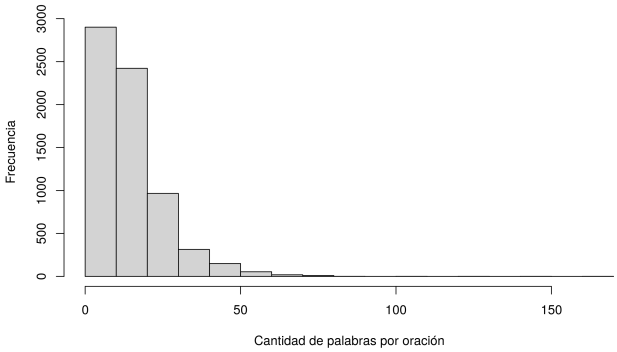
2. Distribuciones en el texto

En esta fase se descompone el texto en oraciones, es decir, se parte después de cada punto. La comparación se realiza a partir de la medición de la cantidad de palabras, y comas que hay en las oraciones de cada texto. Además se analiza la longitud de las palabras usadas.

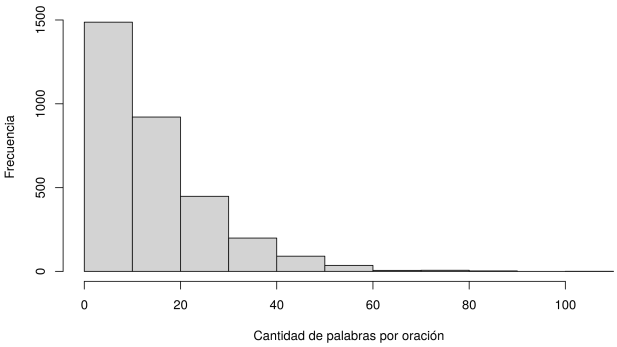
En la figura 1 se puede observar una síntesis de cómo están distribuidos estos elementos en las oraciones de cada texto. En general, tienen un comportamiento similar. Se podría decir que la distribución de la cantidad de palabras por oración en ambos textos (figura 1a y 1b) y la distribución de la cantidad de comas por oración (figura 1c y 1d) es muy similar a una distribución geométrica.

2.1. Experimentos de Bernoulli

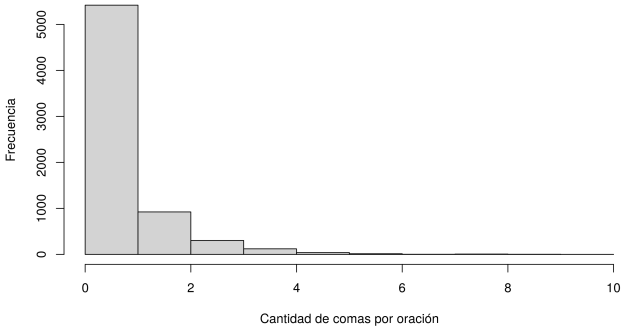
Para esta fase se usa solo el texto de *Anne of Green Gables*. A partir de esa información se realizan distintos experimentos de Bernoulli que buscan imitar el comportamiento de tres distri-



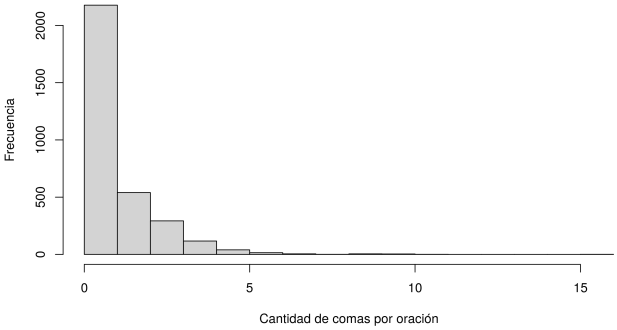
(a) Palabras por oración en *Anne of Green Gables*



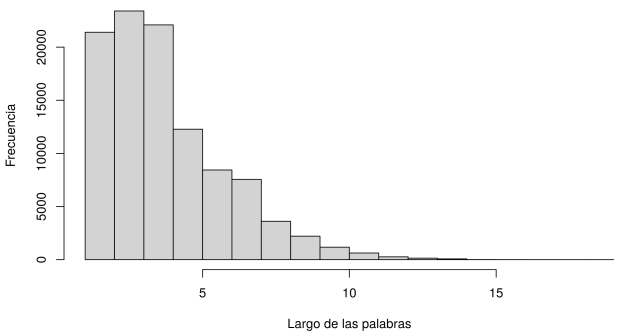
(b) Palabras por oración en *Peter Pan*



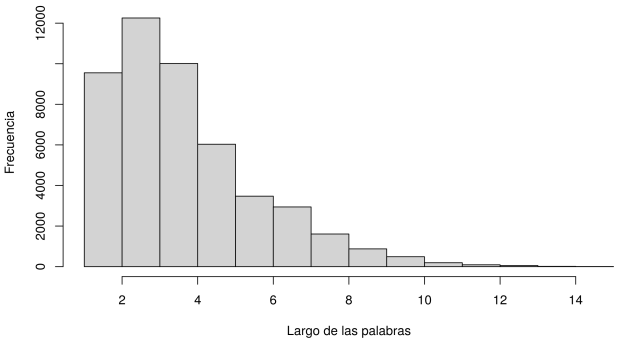
(c) Comas por oración en *Anne of Green Gables*



(d) Comas por oración en *Peter Pan*



(e) Largo de palabras en *Anne of Green Gables*



(f) Largo de palabras en *Peter Pan*

Figura 1: Síntesis de la distribución de las oraciones en los textos.

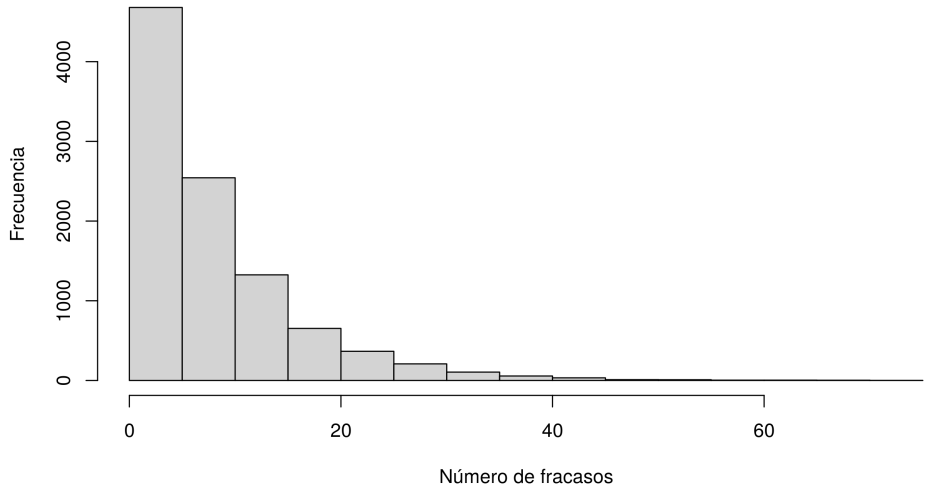


Figura 2: Distribución geométrica.

buciones: geométrica, binomial y binomial negativa.

La distribución geométrica muestra la distribución de la cantidad de repeticiones del experimento que fueron fracasos antes de obtener el primer éxito [5]. Para replicar este comportamiento se define que el éxito se obtiene cuando al elegir al azar una letra del texto esta es una "e", por lo tanto, si se obtiene cualquier otro caracter se considera un fracaso. Al repetir 10,000 veces el experimento se obtienen los resultados de la figura 2. Por la construcción del experimento, es claro que el comportamiento sigue una distribución geométrica.

El siguiente experimento contesta a la pregunta ¿cuántos intentos se necesitan para obtener siete veces el caracter "t", en una elección al azar?. Un éxito, para este experimento, se considera cuando en la elección se obtiene el caracter "t". Esta distribución se conoce como binomial negativa. En la figura 3 se muestran los resultados obtenidos al repetir 10,000 veces el experimento.

La última situación considera un experimento de 30 intentos que se replica 1,000 veces. En cada réplica se cuenta la cantidad de veces que se elige al azar una palabra y ésta tiene una longitud menor o igual a cinco caracteres. La distribución obtenida de este experimento se muestra en la figura 4. La construcción del experimento permite concluir que este comportamiento es el de una distribución binomial.

La distribución en la figura 1e indica que la mayoría de las palabras en el texto tiene una longitud menor o igual a cinco caracteres. De esta forma en el experimento de elegir 30 palabras al azar, se espera que la mayoría de ellas tenga esta restricción en la longitud es por esto que se observa el comportamiento de la figura 4.

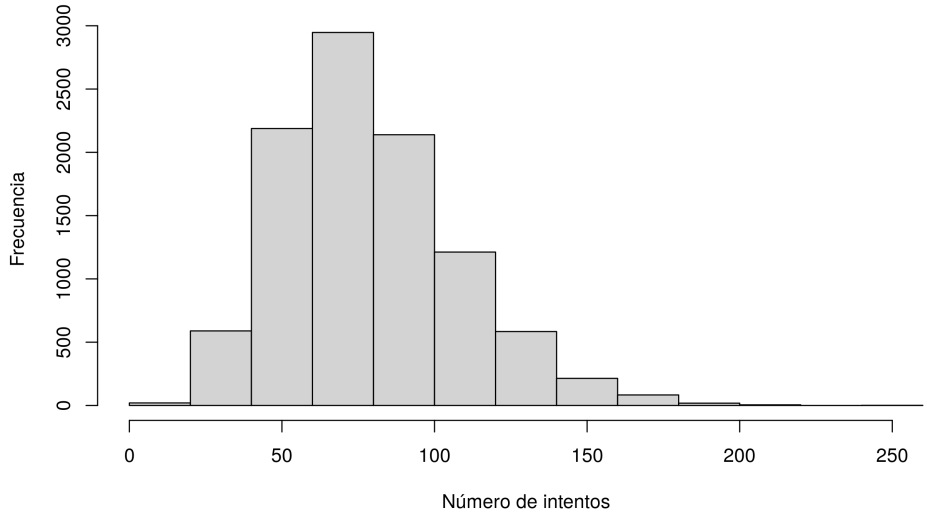


Figura 3: Distribución binomial negativa.

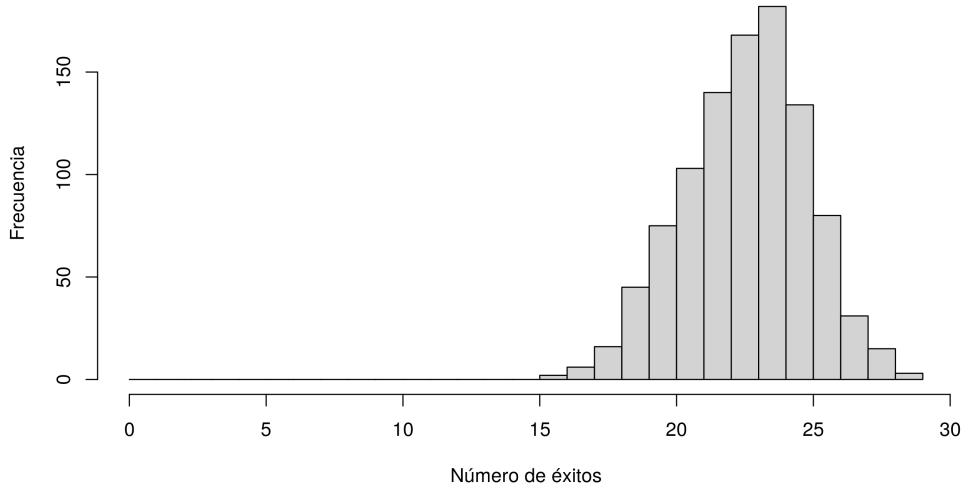


Figura 4: Distribución binomial.

Referencias

[1] James M. Barrie. *Peter Pan (Peter Pan and Wendy)*. Charles Scribner's Sons, 1911.

[2] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/.

[3] Project Gutenberg. Project Gutenberg: library of free eBooks. http://www.gutenberg.org/.

[4] L. M. Montgomery, M. A. Claus, and W. A. J. Claus. *Anne of Green Gables*. L.C. Page & Co, 1908.

[5] Elisa Schaeffer. Modelos probabilistas aplicados: notas del curso. https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html#ut2.

[6] J. Alberto Benavides Vázquez. Modelos probabilistas aplicados – tarea 3. https://github.com/jbenavidesv87/probabilidad/blob/master/tema3/tarea.ipynb.

[7] Gabriela Sánchez Y. Modelos probabilistas aplicados. https://github.com/Saphira3000/MPA.