Ajuste de distribución

Fabiola Vázquez

22 de septiembre de 2020

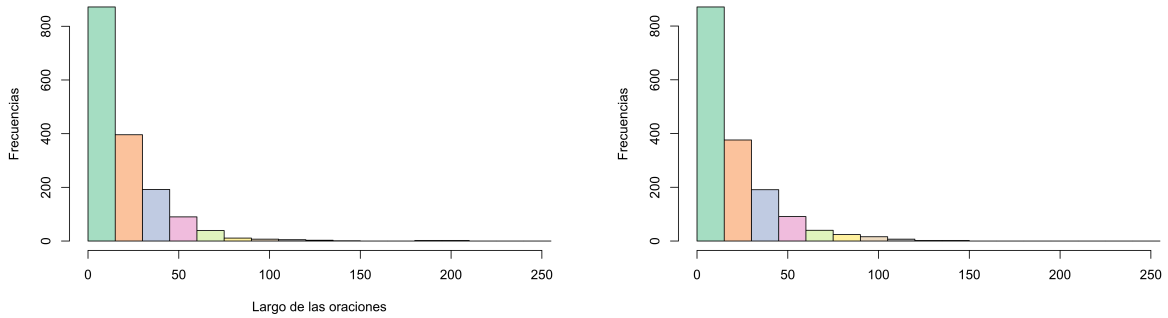
---

1. Introducción

En un trabajo anterior [5] se realizó un estudio de la novela *Alice's Adventures in Wonderland* [1], donde se obtuvieron las palabras y las letras más frecuentes que aparecen en el libro. El objetivo de este estudio, que se realiza con el software estadístico R versión 4.0.2 [3] en el IDE **R Studio** [4], es estudiar las distribuciones de algunos de estos elementos del texto.

2. Análisis

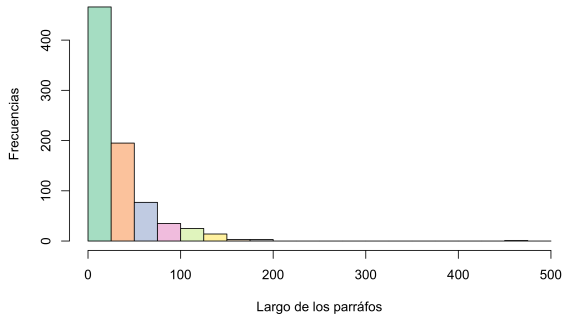
Se calcula el largo de las oraciones en el texto y se realiza un histograma con estos datos, mismo que se puede ver en la figura 1a. Por nuestro conocimiento de distribuciones de probabilidad discretas, suponemos que el largo de las oraciones sigue una distribución geométrica. Con el uso de la función `fitdist` de la librería `fitdistrplus` se ajustan los datos a la distribución geométrica, obteniendo un parámetro `p=0.046`. Se generaron números pseudoaleatorios con esta distribución, tantos como oraciones en el texto, haciendo uso de `rgeom`. Un histograma de estos números se muestra en la figura 1b. La figura 1 muestra ambos histogramas lado a lado.



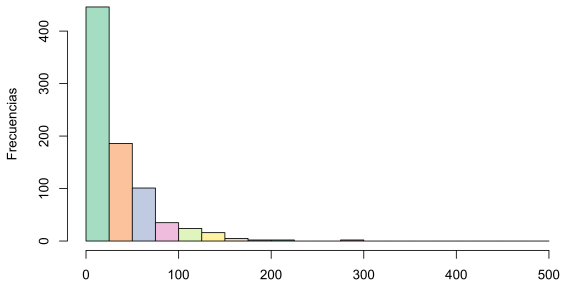
(a) Largo de las oraciones.

(b) Números pseudoaleatorios.

Figura 1: Distribución del largo de oraciones del libro.

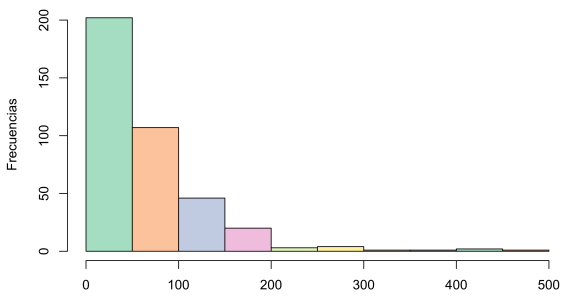


(a) Largo de los párrafos.

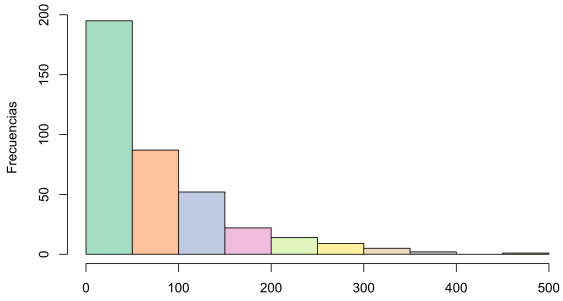


(b) Números pseudoaleatorios.

Figura 2: Distribución del largo de párrafos.



(a) Aparición de la palabra Alice.



(b) Números pseudoaleatorios.

Figura 3: Distribución de la aparación de la palabra *Alice*.

Siguiendo el mismo procedimiento para el largo de los párrafos, se concluye que los datos se ajustan a una distribución geométrica con parámetro `p=0.028`. En la figura 2 se muestra el histograma de las longitudes de los párrafos del libro y el histograma de números pseudoaleatorios generados con distribución geométrica.

Después, se procede a buscar la distribución de la palabra *Alice* en el texto. Para esto, se modela como un experimento Bernoulli, considerando como un éxito la aparición de la palabra *Alice* y un fracaso cualquier otra palabra. Se obtuvo la cantidad de palabras que ocurrían entre cada aparición de la palabra *Alice*, cuyo histograma se muestra en la figura 3a. Se obtiene que en promedio hay aproximadamente 66 palabras entre las apariciones de *Alice*. Esto representa el promedio de fracasos antes de un éxito en una secuencia de experimentos Bernoulli, por lo que debe de seguir una distribución geométrica. Se sabe que esta distribución con parámetro p tiene media $\frac{1-p}{p}$ [2], por lo que al despejar se obtiene que $p = \frac{1}{67}$. El histograma de números psudoaleatorios con distribución geométrica con parámetro `p=1/67` se muestra en la figura 3b.

Referencias

[1] Lewis Carroll. *Alice's Adventures in Wonderland*. Macmillan, Oxford, England, 1865.

[2] George Casella and Roger L. Berger. *Statistical inference*. Thomson Learning, Australia ; Pacific Grove, CA, 2nd edition, 2002.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[4] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Massachusetts, United States, 2020. `http://www.rstudio.com/`.

[5] Fabiola Vázquez. Minería de datos. `https://github.com/fvzqa/Metodos-Probabilisticos/blob/master/Tarea2/Tarea2.pdf`.