

distribucion de frecuencias en el numero de palabras del libro:
Precepts in Practice; or, Stories Illustrating the Proverbs by A. L.
O. E.

TAREA 3

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

En el presente trabajo se hizo el análisis de la frecuencia en el uso de una selección de palabras contenidas en un texto literario con el objeto de calcular las probabilidades de que aparezcan determinadas palabras con cierto número de letras.

El libro que usamos es el de Precepts in Practice; or, Stories Illustrating the Proverbs del autor A. L. O. E. [1] Este libro se encuentra en la biblioteca digital de Gutenberg [2], el cual es el repositorio de más de 63,127 títulos en formato ebook.

Esta obra que revisamos está escrita en inglés y se conforma por un conjunto de relatos cuyas ilustraciones tienen el objetivo de complementar visualmente cada proverbio y, de acuerdo con este género literario, cada historia conlleva una enseñanza de la que podemos aprender en nuestra vida cotidiana, según las diferentes situaciones y asuntos de que se trate.

2. Metodología

Para cumplir con nuestro objetivo, primeramente, procedimos a descargar el texto con el programa “R-4.0.2” [4], dado que la biblioteca Gutenberg [2] permite tal descarga libremente. Como siguiente paso, se removió todo carácter especial no alfa numérico, es decir, los símbolos tipográficos (guiones, cursivas, negritas, etc.) para limpiar el texto. Debido a que la frecuencia de datos era dominada principalmente por artículos, pronombres personales, conjunciones y números, se hizo una hoja de datos en Microsoft Excel [3] con este tipo de palabras. Posteriormente, se le ordenó al programa limpiar el texto de dicho contenido. Con el resto del léxico, se procedió a calcular la frecuencia de las palabras restantes con el programa R para ver las probabilidades de distribución cuantitativa. Finalmente, se llevó a cabo el análisis para realizar los histogramas utilizando las siguientes cuatro funciones de distribución de probabilidad.

1. Distribución geométrica
2. Distribución hypergeométrica
3. Distribución binomial negativa
4. Distribución regular

2.1. Distribución geométrica

Con este modelo pudimos observar la probabilidad de obtener palabras con menos de tres letras. Figure 2.1

2.2. Distribución hypergeométrica

Con esta distribución, que es primordial en el análisis de muestras pequeñas, se obtiene la probabilidad de que aparezcan cinco palabras que sean mayores de dos caracteres en una muestra de 100 palabras. Figure 2.2

2.3. Distribución binomial negativa

Con esta distribución se repite un experimento hasta juntar un número de casos requeridos de éxito. En nuestro trabajo, con este modelo se obtuvo la probabilidad de 15 palabras conformadas por más de 6 caracteres. Figure 2.3

2.4. Distribución regular

Con este modelo se obtiene los resultados de un número determinado de experimentos. En nuestro análisis con la aplicación de dicho modelo se obtuvieron las probabilidades de sacar palabras con más de 7 caracteres. Figure 2.4

Referencias

- [1] A. L. O. E. *Precepts in Practice; or, Stories Illustrating the Proverbs*. 2019.
- [2] Michael Hart. *Project Gutenberg*. <https://www.gutenberg.org/>. 1971.
- [3] Microsoft Office. *Microsoft Excel*. <https://www.microsoft.com/es-mx/microsoft-365/excel>. 2011.
- [4] The R Foundation. *The R Project for Statistical Computing*. <https://www.r-project.org/>. 2019.

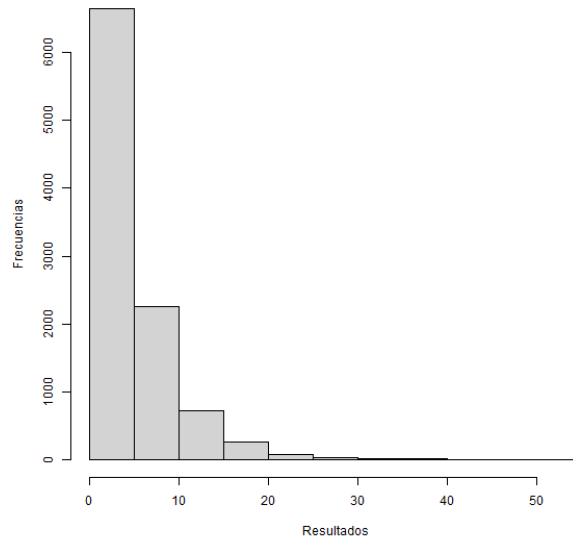


Figura 2.1: Probabilidad de que aparezca una palabra con menos de 3 letras.

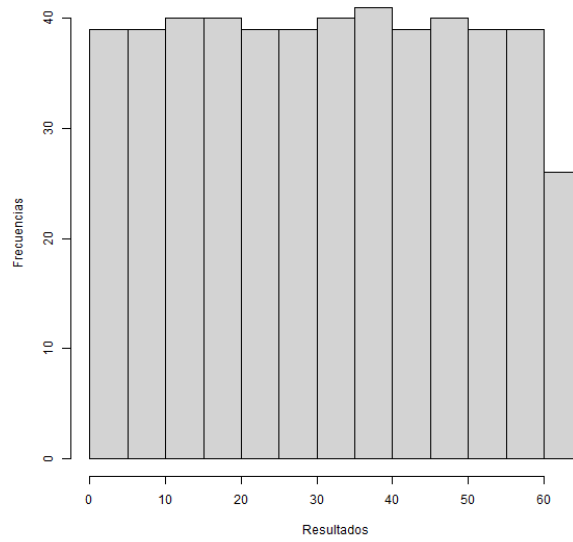


Figura 2.2: Probabilidad de que aparezcan 5 palabras con mas de 2 caracteres.

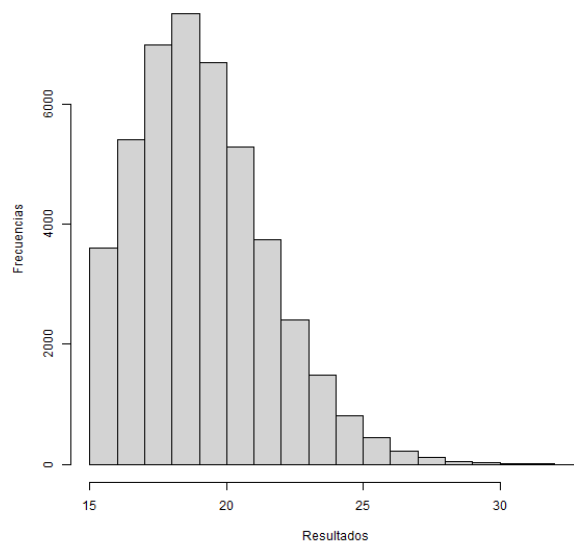


Figura 2.3: Casos de éxito en donde una palabra es mayor que 6 caracteres.

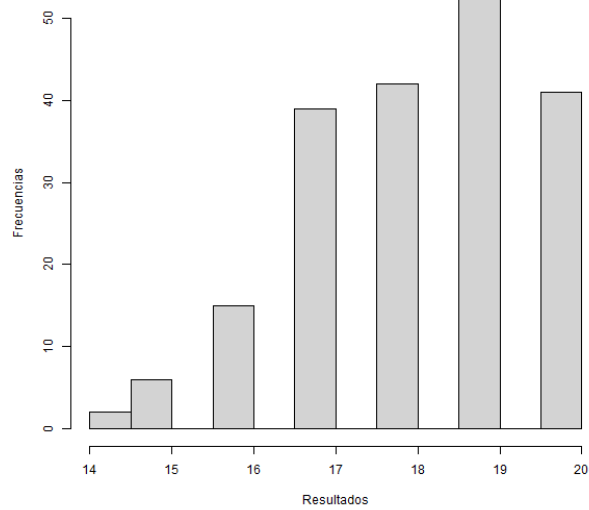


Figura 2.4: Probabilidad de sacar palabras con mas de 7 caracteres.