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

Cambios:

1. secciones,
2. ortografia,
3. referencias a figuras en parentesis.

1. Introducción

En el presente trabajo se busca hacer 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. Este documento se encuentra alojado en el repositorio [4] como recurso libre.

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 [5], 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 normal.

2.1. Distribución geométrica

(Toda esta sección) Si en un experimento existe la posiblidad de obtener éxito con valor p, existe tambien la probabilidad $1-p$ de no tener éxito, asi que deberiamos repetir el experimento hasta obtener un resultado de éxito. La distribución geométrica nos da la probabilidad de que se requieran k repeticiones de ese experimento y se representa por la siguiente función de distribución:

$$P(X = k) = p * (1 - p)^{k-1}.$$

Parámetros	
X	Número de intentos necesarios
p	Probabilidad de éxito
k	Éxito

(1)

La distribución geométrica puede ser representada y graficada en R como se muestra a continuación (Figura 2.1).

```
n      <- 1000
p      <- 0.5
DistribucionGeometrica  <- rgeom(n, p)
hist(DistribucionGeometrica)
```

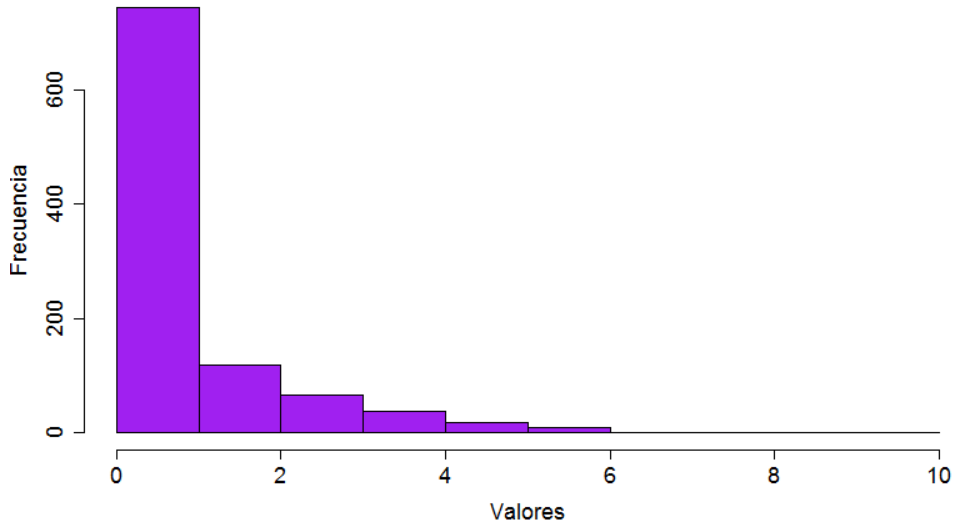

Figura 2.1: Distribución de geométrica con n = 10000.

Aplicando modelo de distribución geométrica, se calcula la probabilidad de obtener palabras con menos de 3 letras en el libro elegido para esta tarea [1], el cual graficado queda de la siguiente forma.

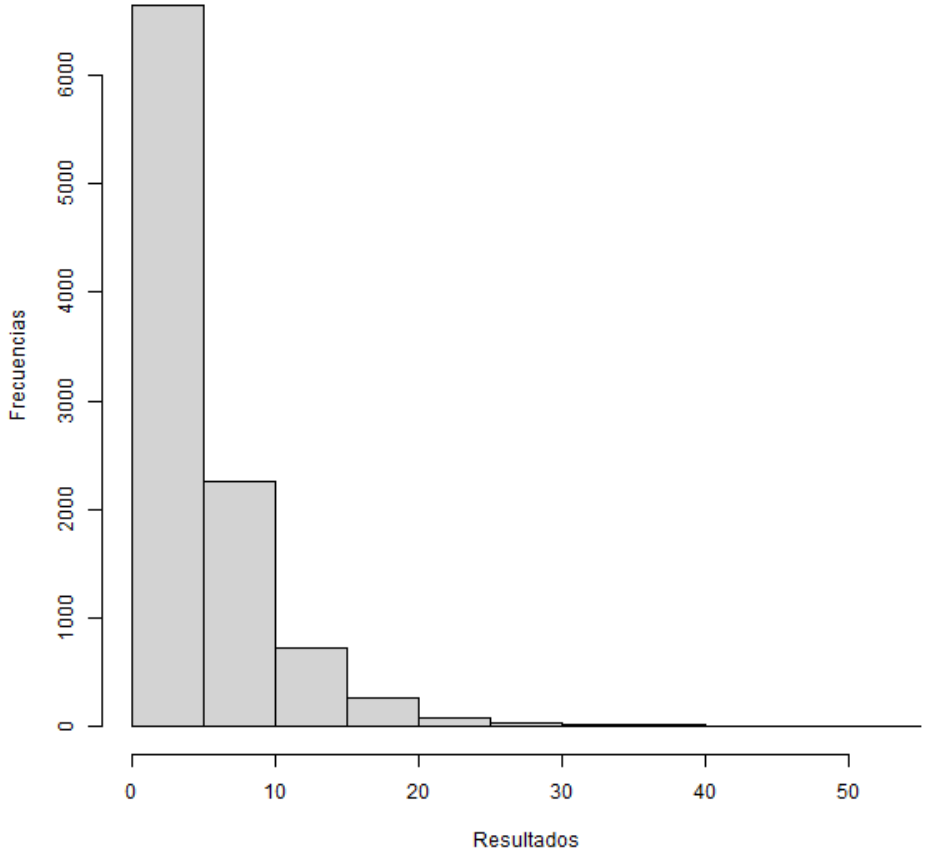

Figura 2.2: Probabilidad de que aparezca una palabra con menos de 3 letras.

2.2. Distribución hypergeométrica

(Toda esta sección) Esta distribución discreta indica la probabilidad de obtener un número de objetos x de una de dos categorías posibles al sacar una muestra de tamaño n sin reemplazo de un total de N objetos, de los cuales k es el tipo requerido.

$$P(X = k) = \frac{\binom{k}{n}\binom{N-k}{n-x}}{\binom{k}{n}}. \tag{2}$$

Parámetros	
N	Tamaño de la población
n	Tamaño de la muestra
k	Cantidad de elementos que cumplen
x	Cantidad que se requiere

La distribución hypergeometrica puede ser representada y graficada en R de la siguiente manera.

```
Size        <- 20
Samples     <- 12
Item        <- 10
```

```
Values        <- 10000
DistribucionHyper  <- rhyper(Values, Size, Samples, Item)
hist(DistribucionHyper)
```

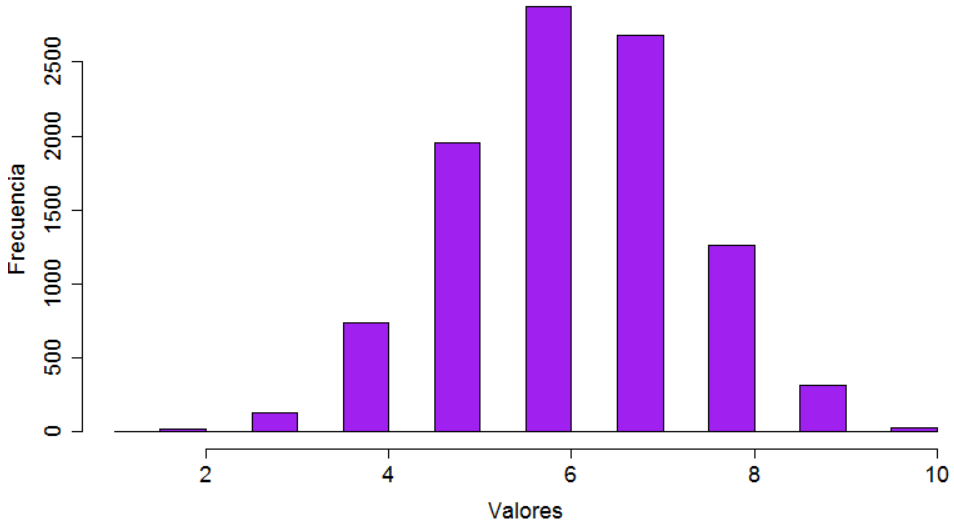

Figura 2.3: Distribución de hypergeometrica con n = 10000.

Con esta distribución, que es primordial en el análisis de muestras pequeñas, se busca la probabilidad de que aparezcan cinco palabras que sean mayores de dos caracteres en una muestra de 100 palabras del libro seleccionado para esta tarea (Figura 2.4).

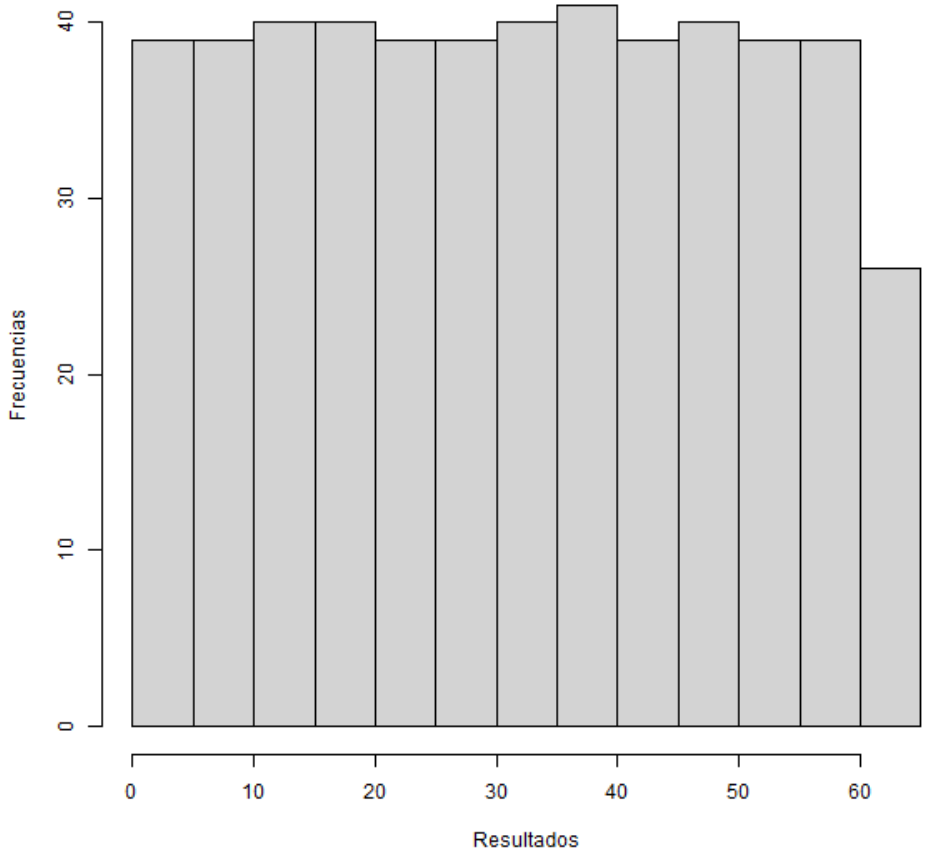

Figura 2.4: Probabilidad de que aparezcan 5 palabras con mas de 2 caracteres.

2.3. Distribución binomial negativa

(Toda esta sección) Es la distribución donde X representa el número de intentos necesarios para alcanzar n éxitos en las realizaciones independientes sucesivas de un experimento de Bernoulli de probabilidad de éxito p. Su función puntual de probabilidad viene dada por:

$$P(X = x) = \binom{n}{x-1} p^n (1-p)^{x-n}.$$

La distribución binomial negativa puede ser representada y graficada en r de la siguiente manera.

```
n      <- 1000
x      <- 10
p      <- 0.25
Distribucionbinomial  <- rnbinom(n, x, p)
hist(Distribucionbinomial)
```

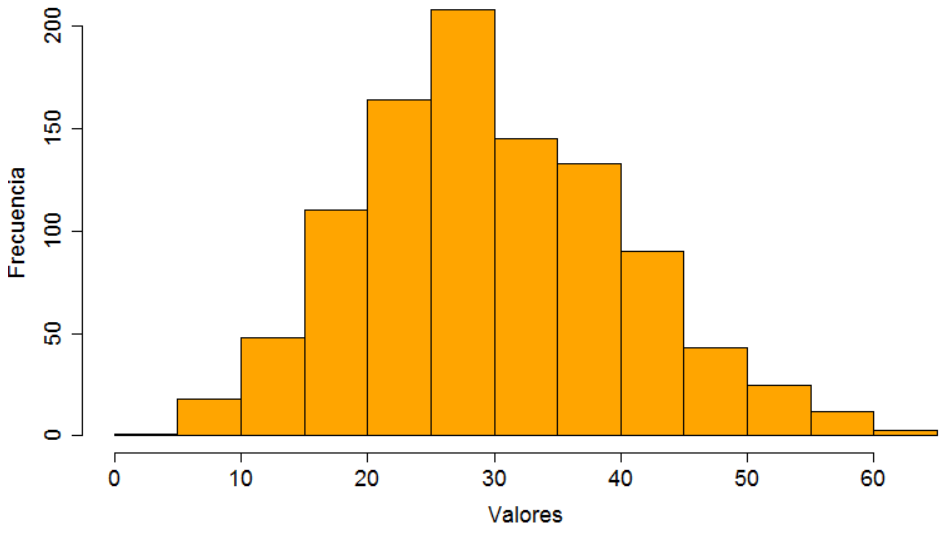

Figura 2.5: Distribución de binomial con n = 10000.

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 (Figura 2.6).

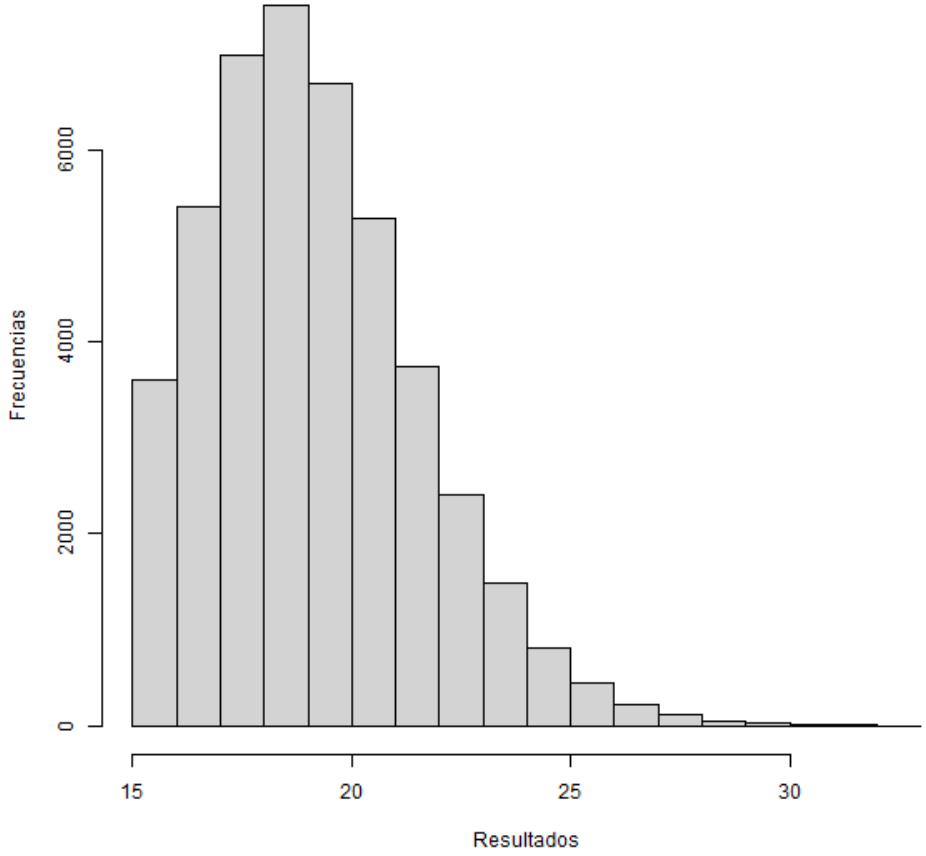

Figura 2.6: Casos de éxito en donde una palabra posee mas de 6 caracteres.

2.4. Distribución normal

(Toda esta sección) La distribución normal es también llamada la distribución gaussiana. Es una distribución continua, una de las más comunes para modelar fenómenos centralizados (estatura y peso). Es la distribución de probabilidad más importante por sus propiedades estadísticas. Supone que en experimentos repetidos, la mayor parte de los resultados coincidirán con un resultado promedio.

$$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(x-\mu)^2}{2\sigma^2}).$$

La distribución normal puede ser representada y graficada en r de la siguiente manera.

```
n      <- 1000
x      <- 23
p      <- 1
Distribucionnormal  <- rnorm(n, x, p)
hist(Distribucionnormal)
```

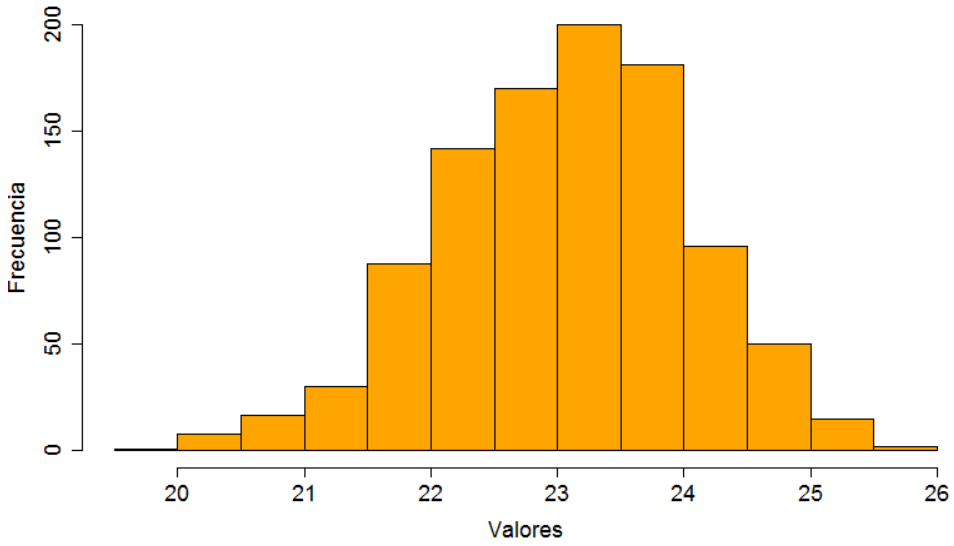

Figura 2.7: Distribución normal con n = 1000 y promedio de x = 23.

Con este modelo se obtiene los resultados de un número determinado de experimentos. Con la aplicación del modelo de distribución normal se quieren obtener las probabilidades de sacar palabras con más de 7 caracteres en el libro elegido para esta investigación (Figura 2.8).

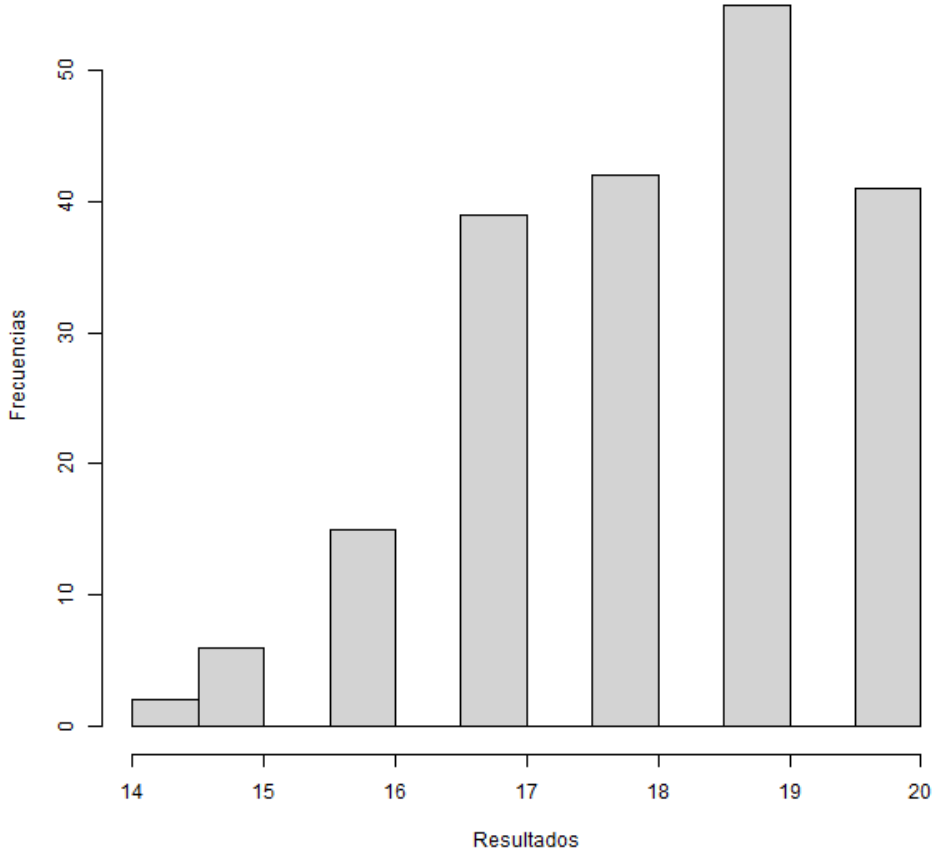

Figura 2.8: Probabilidad de sacar palabras con más de 7 caracteres.

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] Joaquin Arturo Velarde Moreno. *Repositorio con material de la clase de probabilidad*. Recursos libre, disponible en `https://github.com/joaquin3600/Modelos_Probabilistas_Aplicados`. 2020.

[5] The R Foundation. *The R Project for Statistical Computing*. `https://www.r-project.org/`. 2019.