Hoja de Actividades

Matemática y lingüística: La ley de Zipf

Actividad

Lee el siguiente problema:

La base de datos del Corpus de Referencia del Español Actual (CREA) es una herramienta importante, ya que permite acceder a grandes cantidades de datos lingüísticos en un formato estructurado y fácilmente consultable. En esta clase se trabajará con el CREA abreviado con 1000 palabras.

Por ejemplo, a continuación se observan las primeras cinco palabras del CREA:

Ranking	Palabra	Frecuencia absoluta (f)
1	de	9999518
2	la	6277560
3	que	4681839
4	el	4569652
5	en	4234281

¿Se comprueba la ley de Zipf para el CREA?

Responde las siguientes preguntas:

1. Dado el modelo lineal encontrado, ¿cuál sería la frecuencia para la palabra que está en el ranking número 60 y en el ranking número 900?

2. Considerando la base de datos del CREA, ¿cuáles son las frecuencias de las palabras número 60 y 900 del ranking? ¿Cuáles son esas palabras?

3. Considerando el modelo lineal y la base de datos del CREA, ¿cuál es la diferencia entre las frecuencias absolutas para las palabras de la pregunta anterior? ¿Qué se puede decir de esas diferencias?

4. Analiza diferentes rankings (por ejemplo: 1, 11, 25, 500, 800, 1000) y responde: ¿qué tan bien predice las frecuencias el modelo? ¿En qué rangos del ranking funciona mejor el modelo? Justifica.