Descifrando Códigos

Estadística.

Importante:
Escriban el código R para TODAS las preguntas de programación.
Cada grupo debe redactar sus resultados.

Vamos a estudiar algunos algortimos aleatorios que sirven para descifrar mensajes secretos como:

$$\Im\heartsuit\infty\aleph$$

Voy a asumir que cada símbolo corresponde a una letra. Por ejemplo, el mensaje de arriba puede significar:

hola

En ese caso: $\Im \longleftrightarrow h$; $\heartsuit \longleftrightarrow o$; $\infty \longleftrightarrow l$ y $\aleph \longleftrightarrow a$.

El objetivo de los algoritmos que vamos a estudiar es encontrar una correspondencia entre letras y símboos como la de arriba.

Paso 0 Códigos, letras y frecuencias.

Supongo que no se van a sorprender si les digo que en cualquier idioma diferentes letras aparecen con difrentes frecuencias. Por ejemplo la letra mas frecuente en español, inglés y francés es la e.

Pregunta 1 *(4 puntos) Cómo harían para estimar la frecuencia de las letras de un lenguaje? No necesitan escribir un programa en R. Quisiera que describan un algoritmo en pseudo-cćodigo.*

Pregunta 2 *(4 puntos) Escriban un algoritmo que cambie las letras de un string de manera aleatoria. Les recomiendo que primero escojan una codificación de manera aleatoria por ejemplo:* $a \leftrightarrow b$, $c \leftrightarrow z$, *etc. Pueden usar* $sample(letters[14:26], 13)$ *y poner ese vector en correspondencia con* $letters[1:13]$. *Luego, usen la función change*2 *de manera iterativa para cambiar las letras de un texto.*

Pregunta 3 *(4 puntos) Cuál es la probabilidad de una codificación en la que uso* $sample(letters[14:26], 13)$ *y pongo ese vector en correspondencia con* $letters[1:13]$*? Es la misma probabilidad que la de una correspondecia cualquiera?*

Paso 1 Una Primera Idea: Identificar Letras por sus Frecuencias.

Pregunta 4 *(4 puntos) Usen los datos en la tabla frecuencia letras y dibujen un gráfico de barras con las frecuencias para cada letra.*

Pregunta 5 *(4 puntos) Usen su código del paso 0 para codificar tres textos: uno corto (≈ un par de palabras); uno mediano (≈ un par de párrafos) y uno largo (un libro, o el capítulo de un libro o un artículo largo).*

Pregunta 6 *(4 puntos) OK! Hora de descifrar sus primeros mensajes. Traten la siguiente estrategia: Estimen la frecuencia con la que aparecen las letras en sus textos codificados. Luego relacionen las letras por sus frecuencias. Por ejemplo, si en su texto codificado la letra que aparece de manera mas frecuente es la w, entonces supongan que* $w \longleftrightarrow e$.

Su esquema funciona en los tres tipos de códigos?

Paso 2 Markov Chain Monte Carlo: Usar pares de letras.

En esta parte vamos a considerar la frecuencia con la que aparecen 2 letras consecutivas.

Pregunta 7 *(4 puntos) Estimen la frecuencia con la que aperecen todas las posibles combinaciones de dos letras. Cuántas combinaciones posibles de dos letras hay?*

Pregunta 8 *(4 puntos) Programen el siguiente algoritmo donde el score de una correspondecia es*

$$q(thiqq) := P(th) \times P(hi) \times P(iq) \times P(qq) \tag{1}$$

Algorithm 1 MCMC

```
1: procedure MCMC
2:     Escojan una correspondencia de manera aleatoria c.
3:     for do i = 1, ..., n:
4:         Escojan dos letras de la correspondencia al azar.
5:         Cambien su posición en la correspondencia.
6:         Calculen el score q* de la nueva correspondencia c*.
7:         if q* > q then conserven el cambio.
8:         if q > q* then
9:             Conserven el cambio con proba q*/q.
10:            Rechazen con proba (1 − q*)/q.
```

Pregunta 9 *(4 puntos) Funciona el algortimo 1? En qué se diferencia del algoritmo del paso 1?*